DETERMINATION OF THE COMPOSITION OF FAILURE TIME MODELS WITH LONG-TERM SURVIVORS

Abdullah Al Masud

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University
February 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Wanzhu Tu, Ph.D., Co-chair

Zhangsheng Yu, Ph.D., Co-chair

Ying Zhang, Ph.D.

December 8, 2016

Yiqing Song, M.D., Sc.D.

DEDICATION

To my parents, Md. Abdus Samad and Mrs. Hasina Akhter, and to my sister.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisors Dr. Wanzhu Tu and Dr. Zhangsheng Yu for their constant guidance, encouragement and support in my Ph.D. study and dissertation research. I really appreciate the opportunity that they lead me to grow in this wonderful research area. Their guidance has not only trained my knowledge and expertise, but also cultivated me with open-mindedness, ability of critical thinking, and skills of effective communication that are essential for my future career. I have also learned from them the spirit of hard working, persistence, patience and creativity, which I will benefit from for the rest of my life. I would like to specially thank other advisory committee members, Dr. Ying Zhang and Dr. Yiqing Song for their critical evaluations on this dissertation research.

I feel quite grateful to this wonderful Biostatistics Ph.D. program, faculty and staff to provide the friendly and interdisciplinary research environment. I also must thank the classmates and friends who have helped and supported me. I would always remember the joy shared with them.

Finally, I sincerely thank my wife, Rumana Islam (Moon), for her unconditional support, sacrifices, kindness, and optimism to make this journey come true.

Abdullah Al Masud

DETERMINATION OF THE COMPOSITION OF FAILURE TIME MODELS WITH LONG-TERM SURVIVORS

Failure-time data with long-term survivors are frequently encountered in clinical investigations. A standard approach for analyzing such data is to add a logistic regression component to the traditional proportional hazard models for accommodation of the individuals that are not at risk of the event. One such formulation is the cure rate model; other formulations with similar structures are also used in practice. Increased complexity presents a great challenge for determination of the model composition. Importantly, no existing model selection tools are directly applicable for determination of the composition of such models. This dissertation focuses on two key questions concerning the construction of complex survival models with long-term survivors: (1) what independent variables should be included in which modeling components? (2) what functional form should each variable assume? I address these questions by proposing a set of regularized estimation procedures using the Least Absolute Shrinkage and Selection Operators (LASSO). Specifically, I present variable selection and structural discovery procedures for a broad class of survival models with long-term survivors. Selection performance of the proposed methods is evaluated through carefully designed simulation studies.

Wanzhu Tu, Ph.D., Co-chair

Zhangzheng Yu, Ph.D., Co-chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Research Background

In failure-time data analysis, it is assumed that all subjects will eventually experience the underlying disease or event of interest, in the absence of censoring. In some situations, however, a *portion of* the individuals are completely free of the event risk, even after prolonged periods of observation. A time-to-event study in which a substantial portion of subjects having this characteristic, is referred to as a study of "cured" subjects or "long-term survivors". In contrast, non-cured subjects or short-term survivors develop the event during the observation time. The Cox regression model does not directly account for the "cured" portion of the sample, because it assumes all subjects are susceptible to the event (Cox, 1972). One challenge in accommodating the long-term survivors is that they do not experience the event. As a result, event times of the cured subjects are *always* censored.

The cure rate model (or, cure model) is a class of survival models for such data. These models have been applied to a wide range of medical and health science investigations in the last two decades. The models not only quantify the proportion of risk-free individuals, but also estimate the relative risk of the underlying event in those who *are* at risk. A major appeal of this modeling approach lies in its ability to assess the influences of risk factors on the event-time distribution while accounting for factors associated with the cure probability. In this sense, a risk-and-outcome association is depicted by two parameters: (1) the incidence rate of the event among *all* subjects, and (2) the time-to-event distribution for those who experience the event, with and without the risk factor.

Compared to traditional proportional hazard models, the cure models are more difficult to interpret, especially when a large number of independent variables are involved. To alleviate, analysts typically resort to model selection procedures to reduce model complexity. In this dissertation, a model selection procedure is considered as a process in which models are reduced, to achieve desired level of parsimony without sacrificing interpretability. In analytical practice, correct specification of a model underlies the validity of statistical inference.

Along this line, a model selection procedure simultaneously performs two important tasks: (1) identification of the independent variables that are related to the outcome, (2) discovery of the true functional relationship between the independent variables and the survival outcome. For survival models with long-term survivors, such procedures have not been well developed.

The purpose of this dissertation is to present a model selection framework for analysis of survival data with long-term survivors. Variable selection and nonlinear component detection are the key elements of this work. Among other things, the existing general purpose selection procedures do not simultaneously select independent variables and identify hidden functional structures. This dissertation is intended to fill the methodological gap in the context of cure rate models and other models with similar structures.

The dissertation included three interrelated topics, all based on the penalized likelihood method:

1. Variable selection in the mixture and time-promotion cure rate models. Here I assume that all variables have linear effects. I use the L_1 penalty to select the independent variables that are associated with the outcome variables.
2. Nonlinear component detection. For partially linear mixture cure rate models, I additionally select the nonlinear effects by selecting the regression spline functions under the L_1 penalty.

3. Model selection for lifetime data with long-term survivors in the presence of repeated measures. Here I consider a general class of mixture models for analysis of recurrent events data. Structurally, the mixture survival models are similar to the mixture cure rate models.

1.2 Literature Review

In this section, I briefly describe the existing work on analysis of time-to-event data with long-term survivors. Technique details of these methods are delineated in subsequent chapters in contrast with the proposed methods.

The first cure rate models were proposed to analyze time-to-event data with long-term survivors. Earlier studies include Boag (1949), Berkson and Gage (1952), Farewell (1977, 1982). These models are generally referred to as the mixture cure rate models because they have two separate modeling components, one logistic component for the probability of cure, and another component for the depiction of the failure time distribution. Following Farewell's (1977, 1982) mixture formulation, others have extended the cure rate model for various data settings. Notable extensions in the works done by Kuk and Chen (1992), Peng and Dear (2000), and Sy and Taylor (2000). More recently, Lambert *et al.* (2010) applied the proportional hazards assumption to cure rate models. Chen and Wang (2000) and Zhang and Peng (2009) extended the basic cure model structure to accelerated hazards model settings.

An alternative method, the promotion time cure rate model has been proposed by Yakovlev *et al.* (1996) and Tsodikov (1998). This alternative model was initially derived from kinetic studies of carcinogenic cells, where the numbers of carcinogenic cells were considered as the underlying promotors of cancer recurrence and modeled by Poisson regression, while time to cancer recurrence was modeled by proportional hazard models. Chen *et al.* (1999), Chen and Ibrahim (2001), and Yin and Ibrahim (2005) extended this model to a Bayesian framework.

Mixture cure rate model and promotion time cure rate model are both frequently used in medical investigations, albeit with somewhat different motivations. A common challenge in analyzing survival data with long-term survivors, however, is to identify the variables that are truly relevant to the event of interest. The challenge becomes more severe with increased model complexity and growing number of variables (Breiman, 1996). One approach to overcome the challenge is to introduce a penalty that forces out the less relevant variables. In linear and Cox models, one such penalty is Tibshirani's (1996, 1997) Least Absolute Shrinkage and Selection Operator (LASSO). In Cox models, Fan and Li (2002) later proposed the use of Smoothly Clipped Absolute Deviation (SCAD) penalty, and Zhang and Lu (2007) used an adaptive LASSO procedure.

Model selection procedures have been developed various modeling settings. Among many others, Lin and Zhang (2006) proposed a model selection procedure in nonparametric regression models but their method does not determine the functional forms of the variables. Li and Liang (2008) proposed selection method to select functional effects of a nonlinear variables under a framework of multiple hypotheses testing. But this method often suffers from failure to correctly adjusting the type I error rate, especially when the model includes a large number of functional variables. Zhang *et al.* (2011) proposed an automatic structural discovery procedure for partially linear models and thus generalizing Lin and Zhang (2006)'s work. Partial linear model has combination of linear and nonparametric effects of variables of the data. These developments have made it easier for researchers to select variables through data, and to construct models with appropriate structures. To accommodate the nonlinear effects, various spline techniques can be used (O'Sullivan, 1988; Eilers and Marx, 1996). Discovery of the nonlinear effects requires decomposing a nonlinear variable into the linear parts and the nonlinear parts. Wand and Ormerod (2008) illustrated the idea with cubic B-spline function using a spectral decomposition technique. In application

of survival data, Lin and Halabi (2013) investigated the structure of the Cox model and estimated the risk factors with partially linear Cox model. My objective is to develop flexible methods for model specification and variable selection. Expectation-maximization (EM) algorithm is derived and performed for structural discovery and model estimation. Inferences are conducted based on the bootstrap procedure.

Mixture cure rate models have been extended to repeatedly measured failure time data. Studies such as Yau and Ng (2001), Yu (2008), Lai and Yau (2008), Li *et al.* (2010), Rondeau (2010), and Rondeau *et al.* (2011) have provided good examples of such extensions. More recently, Yan and Huang (2012) added time-varying coefficients to the selection procedure. But variable selection and functional form determination for mixture cure rate models have not been developed for repeated failure time data. I intend to fill in this knowledge gap.

1.3 Organization of Dissertation

This dissertation has three methodological components: (1) Variable selection in mixture cure and promotion time cure rate models. Full methodological details are described in Chapter 2, and is reported in paper to appear in *Journal of Statistical Methods in Medical Research*. (2) Structural discovery in cure rate models. The method is described in Chapter 3. (3) Model selection and structural discovery in survival analysis with long-term survivors and repeated measurements. The details are presented in Chapter 4.

Chapter 2

VARIABLE SELECTION FOR MIXTURE AND PROMOTION TIME CURE RATE MODELS

This chapter presents two choices of cure models (or cure rate models) for analysis of time-to-event data from subjects when a certain portion of them will not develop the underlying event although long they are followed. It starts with a short description of the motivating research question about onset of wheezing symptom in children and the rational behind the proposed method for variable selection. The formulation of the method is described in detail, followed by an analysis of childhood wheezing data that demonstrates how the proposed method offers to analyze a varieties of complex survival data.

2.1 Research Background

Standard survival models, such as the frequently used Cox regression models, assume that all subjects are susceptible to the event of interest, and that all subjects will eventually experience the outcome if the follow-up is long enough (Cox, 1972). Data from some applications, however, contradict the notion that *all* subjects are at risk. In practice, analysts deal with the situation by treating the risk-free subjects as "cured". Compared to the non-cured, the cured tend to have much extended survival times, as indicated by long flat tails and heavy right censoring in Kaplan-Meier curves (Sy and Taylor, 2000).

Data with such characteristics are abundant in clinical studies. For instance, childhood wheezing, an airway symptom defined by a coarse or whistling breathing sound, tends to occur only in certain children, while others never exhibit wheezing

symptoms in early years of life (Tepper *et al.*, 2008). Data from the study showed that Kaplan-Meier curves of the onset age of wheezing essentially flattened after the first 48 and 32 months of life in girls and boys, and thus confirming the existence of risk-free subgroups. See Figure 2.1. Data with similar features are also seen in immuno-oncological studies (Chen, 2013).

Cure rate models are standard techniques for such data. Traditional cure rate models assume that the population consists of both cured and non-cured subjects (Boag, 1949). The standard formulation is a mixture of logistic regression and survival analysis, with the former quantifying the cured portion and the latter depicting the event time distribution of the non-cured (Farewell, 1982). This mixture has been the basis of several model extensions (Taylor, 1995; Sy and Taylor, 2000; Peng and Dear, 2000). A more biology-motivated approach is the promotion time cure model, proposed by Yakovlev and colleagues (1993) in the context of cancer recurrence. Briefly, Yakovlev's model assumes that cancer recurrence is promoted by carcinogenic cells that remain active after treatment. So the unobserved number of carcinogenic cells is incorporated into the analysis through a Poisson model. This line of models has been further extended by others, mostly in the Bayesian framework (Chen *et al.*, 1999, 2002; Ibrahim *et al.*, 2002; Tsodikov *et al.*, 2003). The two different modeling approaches have been compared by a number of authors (Broët *et al.*, 2001; Yin and Ibrahim, 2005).

Regardless of one's modeling preference, a common challenge faced by analysts is to select the right independent variables for the intended model. With the complex structures of cure rate models, variable selection is certainly not a trivial exercise. Among other things, traditional stepwise procedures often lack the desired stability (Breimman, 1996). Following Tibshirani's works on the Least Absolute Shrinkage and Selection Operators (LASSO) (Tibshirani, 1996, 1997), penalize likelihood-based regularization methods have been developed for variable selection in frequently used

statistical models, including the traditional Cox regression models (Cox, 1972). Theoretically, some of these methods have been shown to possess the oracle properties (Fan and Li, 2001; Zou, 2006; Zhang and Lu, 2007). Most recently, attempts have been made to extend the LASSO-based selection approach to joint models of longitudinal and survival outcomes (He *et al.*, 2015). The successful use of LASSO in complex models points to the plausibility of a similar application in the cure rate models.

Literature on variable selection in cure rate models is relatively sparse. One notable piece of work in this field is by Liu and colleagues (2012) who proposed to use LASSO with a Smoothly Clipped Absolute Deviation (SCAD) penalty to select variables for the mixture cure rate model. The non-convex form of the SCAD penalty, however, tends to increase the difficulty of parameter estimation. As a result, estimators often lack numerical stability (Zhang and Lu, 2007). Alternatively, Zou (2006) proposed an adaptive LASSO method with L_1 penalty, which is computationally more stable in comparison with SCAD.

In this research, I discuss variable selection in mixture and promotion time cure models using LASSO and adaptive LASSO. To the best of my knowledge, this is the first study of its kind, especially for the promotion time cure model. I compare the selection performance of LASSO and adaptive LASSO. The methods are easily implementable using an expectation-maximization (EM) algorithm, with generally consistent performance. An extensive simulation study is conducted to evaluate the operational characteristics of the procedures in both modeling settings. Finally, I apply the methods to select variables for a mixture cure model using data from a study of childhood wheezing.

2.2 Models and Estimation

2.2.1 Mixture Cure Rate Models

Model:

Let $\tilde{T}_i$ and C_i be the respective failure time and censoring time for the ith subject, $i = 1, 2, ..., n$. The observed time is $T_i = \min(\tilde{T}_i, C_i)$. I assume that the censoring time C_i is random and noninformative. I define the failure time indicator as $\delta_i = 1$ if $\tilde{T}_i \leq C_i$ (T_i is observed), and $\delta_i = 0$ otherwise. Let $Y_i = 1$ be a binary indicator for the non-cured, and $P(Y_i = 1) = \theta(\cdot)$. I write the independent variable vectors for the logistic and survival components as $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{z}_i \in \mathbb{R}^q$ respectively; and vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ may share common elements. Under such a notation, the population survival function $S_p(t)$ can be written as

$$S_p(t) = \{1 - \theta(\mathbf{x}_i)\} + \theta(\mathbf{x}_i) S_{nc}(t|\mathbf{z}_i), \tag{2.1}$$

where $S_{nc}(t|\mathbf{z}_i)$ is the survival function of the non-cured, given $\mathbf{z}_i$. As t increases, $S_p(t) \to \{1 - \theta(\mathbf{x}_i)\} > 0$. I note that $S_p(.)$ may not be a proper survival function.

With a logit link function in the mixture cure rate model, Farewell (1982) described the effects of independent variables $\mathbf{x}$ on the probability of not being cured as

$$\theta(\mathbf{x}_i) = \frac{\exp(\beta^T \mathbf{x}_i)}{1 + exp(\beta^T \mathbf{x}_i)},$$

where β is a vector of regression coefficients for $\mathbf{x}_i$. For the non-cured, the Cox proportional hazards (PH) model can be written as $\lambda_{nc}(t|\mathbf{z}_i) = \lambda_{nc,0}(t) e^{\gamma^T \mathbf{z}_i}$, where γ is the coefficient vectors for $\mathbf{z}_i$, and $\lambda_{nc,0}(t)$ is the baseline hazard. The cumulative baseline hazard function is $\Lambda_{nc,0}(t) = \int_0^t \lambda_{nc,0}(u) du$. The independent variable effects

for the non-cured in Model (2.1) are interpreted in a way similar to that in the traditional Cox models.

Variable Selection and Estimation

For simplicity, I denote the observed data from the ith subject as $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$. The likelihood function of model(2.1) is

$$L(\beta, \gamma, \lambda_{nc,0}) = \prod_{i=1}^{n} \{\theta(\mathbf{x}_i)\lambda_{nc,0}(t_i)e^{\gamma^T \mathbf{z}_i} S_{nc,0}(t_i)^{e^{\gamma^T \mathbf{z}_i}}\}^{\delta_i} \prod_{i=1}^{n} \{1-\theta(\mathbf{x}_i)(1-S_{nc,0}(t_i)^{e^{\gamma^T \mathbf{z}_i}})\}^{1-\delta_i}. \tag{2.2}$$

Estimation of the nonparametric baseline hazard is needed to maximize (2.2). Here I use an EM algorithm to maximize the complete likelihood based on $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i, y_i)$, by treating y_i as a latent binary variable. The complete likelihood includes a logistic component for the cured, and a PH component for the non-cured. I write

$$L_C(\beta, \gamma, \lambda_{nc,0}; y) = \prod_{i=1}^{n} \left[\theta(\mathbf{x}_i)^{y_i}(1-\theta(\mathbf{x}_i))^{1-y_i}\right] \prod_{i=1}^{n} \left[\{\lambda_{nc,0}(t_i)exp(\gamma^T \mathbf{z}_i)\}^{\delta_i} S_{nc,0}(t_i)^{e^{\gamma^T \mathbf{z}_i}}\right]^{y_i}.$$

The log-likelihood is

$$l_C(\beta, \gamma, \lambda_{uc,0}; y) = l_{C,1}(\beta; y) + l_{C,2}(\gamma, \lambda_{nc,0}; y) \tag{2.3}$$

For simplicity, I write the first term of the above equation as $l_{C,1}(\beta; y) = \sum_{i=1}^{n} \Big\{ y_i \beta^T \mathbf{x}_i - log\{1+exp(\beta^T \mathbf{x}_i)\} \Big\}$, and second term as $l_{C,2}(\gamma, \lambda_{nc,0}; y) = \sum_{i=1}^{n} \Big\{ y_i \delta_i \{log\lambda_{nc,0}(t_i) + \gamma^T \mathbf{z}_i\} + y_i exp(\gamma^T \mathbf{z}_i) log S_{nc,0}(t_i) \Big\}$. To allow for sparse estimation, I use an adaptive LASSO and impose an L_1 norm penalty on the log likelihood:

$$pl_c(\beta, \gamma, \lambda_{nc,0}; y) = \Big\{ l_{C,1}(\beta; y) - \tau_1 \sum_{j=1}^{p} \frac{|\beta_j|}{|\rho_{1,j}|} \Big\} + \Big\{ l_{C,2}(\gamma, \lambda_{nc,0}; y) - \tau_2 \sum_{k=1}^{q} \frac{|\gamma_k|}{|\rho_{2,k}|} \Big\}, \tag{2.4}$$

where $\rho_{1,j}$ and $\rho_{2,k}$ are the weight parameters, and τ_1 and τ_2 are the tuning parameters controlling the amount of penalty. Values of the tuning parameters can be determined either by cross-validation or by the Bayesian Information Criteria (BIC). I discuss the selection of tuning parameters later in the section.

Following Zou (2006), I use consistent estimators of (β, γ) as the weight parameters (ρ_1, ρ_2). The closer the true estimate to 0, the greater the penalty. As a result, factors with smaller coefficients are more likely to be excluded from the model. The adaptive LASSO essentially shrinks the less important effects to zeros, and thus achieving a more parsimonious model. When ρ_1 and ρ_2 are set to the 1, the method leads to LASSO estimators proposed by Liu *et al.* (2012). In this research, I estimate ρ_1 and ρ_2 by maximizing (2.3).

Computation

For computation, I use adaptive LASSO estimates $(\hat{\beta}, \hat{\gamma})$ and the quadratic approximation algorithm (Fan and Li, 2001).

E-step: Let $(\beta^{(m)}, \gamma^{(m)}, \lambda_{nc,0}^{(m)})$ be the parameter estimates in the mth iteration. Given the observed data $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, in the $(m+1)$th iteration, I replace y_i in (2.3) with $y_i^{(m+1)}$

$$y_i^{(m+1)} = \delta_i + (1-\delta_i)\frac{\theta(\mathbf{x}_i)^{(m)} S_{nc,0}^{(m)}(t_i)^{e^{\gamma^{(m)T}\mathbf{z}_i}}}{1-\theta(\mathbf{x}_i)^{(m)}\{1-S_{nc,0}^{(m)}(t_i)^{e^{\gamma^{(m)T}\mathbf{z}_i}}\}}.$$

M-step: With $y^{(m+1)}$ plugged in, I maximize (2.4) with respect to $(\beta, \gamma, \lambda_{nc,0})$.

The M-step involves the following sub-steps

1. Estimate the cumulative baseline hazard function $\Lambda_{nc,0}(t)$ using a Breslow type estimator (Klein, 1982). Specifically, the nonparametric estimate for

the $(m+1)$th iteration is

$$\Lambda_{nc,0}^{(m+1)}(t) = \sum_{t_l \le t} \frac{d_l}{\sum_{k^* \in R_l} y_{k^*}^{(m+1)} \exp(\gamma^{(m)T} \mathbf{z}_{k^*})},$$

where d_l is the number of events at the earliest time point t_l, and R_l is the number of individuals at risk at t_l.

2. Solve the penalized score equation for $\beta^{(m+1)}$ in the logistic model

$$\begin{aligned} 0 &= U(\beta) = \sum_{i=1}^{n} \Big[y_i^{(m+1)} - \frac{exp(\beta^T \mathbf{x}_i)}{1 + exp(\beta^T \mathbf{x}_i)} \Big] \mathbf{x}_i^T - \tau_1 \sum_{j=1}^{p} \frac{\beta_j / |\beta_j^{(m)}|}{|\rho_{1,j}|} \\ &= \nabla l_{C,1}(\beta; y^{(m+1)}) - \tau_1 \beta^T \psi(\beta^{(m)}, \rho_1), \end{aligned}$$

where $\psi(\beta^{(m)}, \rho_1) = diag\{\frac{1/|\beta_j^{(m)}|}{|\rho_{1,j}|}\}, j = 1, 2, ...p$, and $\nabla l_{C,1}(\beta; y^{(m+1)}) = \frac{\partial}{\partial \beta} l_{C,1}(\beta; y^{(m+1)})$. I obtained the penalty term $\sum_{j=1}^{p} \frac{\beta_j / |\beta_j^{(m)}|}{|\rho_{1,j}|}$ by using a quadratic approximation of the penalized likelihood. The penalized Hessian matrix H_β for β is given by $H_\beta = \frac{\partial}{\partial \beta} U(\beta)$.

3. Solve the penalized score equation for the survival model with respect to $\gamma^{(m+1)}$ with given $\Lambda_{nc,0}^{(m+1)}(t)$, $\beta^{(m+1)}$

$$\begin{aligned} 0 &= U(\gamma) = \sum_{i=1}^{n} \Big[y_i^{(m+1)} \delta_i - y_i^{(m+1)} exp(\gamma^T \mathbf{z}_i) \Lambda_{nc,0}^{(m+1)}(t_i) \Big] \mathbf{z}_i^T - \tau_2 \sum_{k=1}^{q} \frac{\gamma_k / |\gamma_k^{(m)}|}{|\rho_{2,k}|} \\ &= \nabla l_{C,2}(\gamma; \lambda_{nc,0}^{(m)}, y^{(m+1)}) - \tau_2 \gamma^T \psi(\gamma^{(m)}, \rho_2), \end{aligned}$$

where $\psi(\gamma^{(m)}, \rho_2) = diag\{\frac{|1/\gamma_k^{(m)}|}{|\rho_{2,k}|}\}, k = 1, 2, ...q$, and $\nabla l_{C,2}(\gamma; \lambda_{nc,0}^{(m)}, y^{(m+1)}) = \frac{\partial}{\partial \gamma} l_{C,2}(\gamma \lambda_{nc,0}^{(m)}, y^{(m+1)})$. I obtained the penalty term $\sum_{k=1}^{q} \frac{\gamma_k / |\gamma_k^{(m)}|}{|\rho_{2,k}|}$ by using a quadratic approximation for the penalized likelihood. The penalized Hessian matrix H_γ for γ in the $(m+1)$th iteration is obtained by $H_\gamma = \frac{\partial}{\partial \gamma} U(\gamma)$.

The M-step iterates through the above sub-steps until convergence is achieved. The final maximum likelihood (ML) estimates $(\hat{\beta}, \hat{\gamma})$ are achieved by iterating between the E and the M steps.

Alternatively, one could use a parametric function for the baseline hazard $\lambda_{nc,0}$ in (2.3) to simplify process. For a finite partition of follow-up time intervals $0 < s_1 < s_2 < ... < s_G$ with $s_G > \max\{t_i : i = 1, 2, ...n\}$ for a prespecified G, one could assume a constant hazard rate $\lambda_{nc,0}(t) = \alpha_g$ for the gth interval. The estimate $\alpha^{(m+1)}$ of α is obtained by maximizing $l_{C,2}(.)$ with respect to α. For $g = 1, 2, \cdots, G$, it is easy to show that $\alpha_g^{(m+1)} = \left[\sum_{s_{g-1}<t_i\leq s_g} \delta_i y_i^{(m+1)}\right] \times \left[\{\sum_{s_{g-1}<t_i\leq s_g} y_i^{(m+1)}(t_i - s_{g-1}) + \sum_{y_i>s_g} y_i^{(m+1)}(s_g - s_{g-1})\} exp(\gamma^{(m+1)}\mathbf{z}_i)\right]^{-1}$. I later evaluate the selection performance of the nonparametric and parametric baseline hazard function in my simulation study.

In summary, the key steps of the EM algorithms are:

Step 1: Fix the tuning parameter $\tau = (\tau_1, \tau_2)$ and initialize $(\beta^{(0)}, \gamma^{(0)}, \lambda_{nc,0}^{(0)}(t))$

Step 2: Execute the E-step and compute $\lambda_{nc,0}(t)$

Step 3: Update the estimates as $\beta^{(1)} = \beta^{(0)} - H^{-1}(\beta^{(0)})U(\beta^{(0)})$ for logistic regression and

$\gamma^{(1)} = \gamma^{(0)} - H^{-1}(\gamma^{(0)})U(\gamma^{(0)})$ for survival model

Step 4: Repeat step 2 and 3 until $|\beta^{(1)} - \beta^{(0)}| \to 0$ and $|\gamma^{(1)} - \gamma^{(0)}| \to 0$

Regularization/Tuning Parameter Selection

Choosing appropriate tuning parameters $\tau = (\tau_1, \tau_2)$ is essential for variable selection. As τ increases, more coefficients shrink to zero (Zou, 2006). At the same time, estimates of non-zero coefficients are likely to have increased biases (Zhang and Lu, 2007). Nishii (1984) adopted a generalized information criterion (GIC) to select τ. The GIC type regularization parameter selector takes the form

$$GIC(\tau) = \frac{1}{n}\{l_C + \kappa df_\tau\}, \tag{2.5}$$

where df_τ is the degree of freedom associated with Model (2.3). I select the combination of τ_1, τ_2 that minimizes Equation (2.5) for a given κ. As κ increases, the

size of selected model decreases. When $\kappa = \log(n)$, the GIC-type selector reduces to the traditional Bayesian information criterion (BIC) selector (Schwarz, 1978). To solve for β and γ, I use the BIC regularization parameter selector. The BIC selector has been shown to identify the true model consistently (Zou and Li, 2008), and is asymptotic efficient (Zhang *et al.*, 2010).

Post-Selection Inference

Making valid inference in the selected models poses a new set of challenges, which are beyond the scope of the current paper. First, LASSO penalty could introduce biases to parameter estimation. An obvious way to minimize the bias is to fit the selected model without the penalty term. Such a two-stage approach is consistent with the current practice where inferences are based on the selected models, as advocated by standard textbooks (Moore and McCabe, 2009). What left unsaid is the conditional nature of the inference. The validity of such inference is clearly contingent upon the goodness of the selected model. Recently, Berk and colleagues prescribed an attractive solution (Berk *et al.*, 2013). They argued that in linear models, one could treat the post-selection inference as one in a multiple comparison situation, by properly accounting for the errors associated with *all* possible sub-models. While the idea is intuitively appealing, its validity in nonlinear models remains to be validated.

Another issue that affects the inference is the estimation of standard errors of the model parameters. Traditionally, asymptotic standard errors are derived from the Hessian matrix of the observed likelihood. With the use of EM algorithm for estimation, one could simply plug in the model parameter estimates $(\hat{\beta}, \hat{\gamma})$ into the Hessian matrix. The following formulae are typically used to approximate the covariance estimators of $\hat{\beta}$ and $\hat{\gamma}$, respectively, given $\hat{\Lambda}_{nc,0}(t_i)$:

$$H(\hat{\beta}) = -\mathbf{x}_i\{\frac{e^{\hat{\beta}^T\mathbf{x}_i}}{(1+e^{\hat{\beta}^T\mathbf{x}_i})^2}\}\mathbf{x}_i^T + \mathbf{x}_i\{(1-\delta_i)\times\frac{e^{\hat{\beta}^T\mathbf{x}_i}exp\{-\hat{\Lambda}_{nc,0}exp(\hat{\gamma}^T\mathbf{z}_i)\}}{(1+e^{\hat{\beta}^T\mathbf{x}_i}exp\{-\hat{\Lambda}_{nc,0}(t_i)exp(\gamma^T\mathbf{z}_i)\})^2}\}\mathbf{x}_i^T;$$

$$H(\hat{\gamma}) = -\mathbf{z}_i\delta_i e^{\hat{\gamma}^T\mathbf{z}_i}\hat{\Lambda}_{nc,0}(t_i)\mathbf{z}_i^T - \mathbf{z}_i(1-\delta_i) \times \frac{e^{\hat{\beta}^T\mathbf{x}_i} exp\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}\}\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}}{(1+e^{\hat{\beta}^T\mathbf{x}_i} exp\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}\})^2}\mathbf{z}_i^T$$
$$- \mathbf{z}_i\{(1-\delta_i) \times \frac{e^{\hat{\beta}^T\mathbf{x}_i} exp\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}\}\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}}{1+e^{\hat{\beta}^T\mathbf{x}_i} exp\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}\}}$$
$$\times (1 - e^{\hat{\beta}^T\mathbf{x}_i} exp\{-\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i}\}\hat{\Lambda}_{nc,0}(t_i)e^{\hat{\gamma}^T\mathbf{z}_i})\}\mathbf{z}_i^T.$$

Alternatively, one could resort to resampling methods to ascertain the standard error estimates. An advantage of bootstrap standard error estimates is their non-reliance of distributional assumptions. To implement, I resample the observations for a finite number of times with replacement. The resamples are all of size n, the size of the original sample. I then estimate the parameters for each of the bootstrap samples; bootstrap standard errors are calculated from the parameter estimates. With the use of EM algorithm, I use this resampling procedure to obtain the appropriate standard errors of $\hat{\beta}$, and $\hat{\gamma}$.

2.2.2 Promotion Time Cure Model

Development of the selection method for promotion time cure models parallels that of the mixture cure rate models.

Model

Promotion time cure rate model was developed in the context of cancer recurrence led by carcinogenic cells. For example, Chen *et al.* (1999) assume that the ith subject has Y_i carcinogenic cells that could lead to a recurrent disease. They further assume that Y_i follows a Poisson distribution with mean function $\theta(\mathbf{x}_i) = exp(\beta^T\mathbf{x}_i)$, where β is the coefficient vector for independent variables $\mathbf{x}_i \in \mathbb{R}^p$, and that for each cell, time to event ζ follows a distribution $F_1(t)$, or a survival function $S_1(t) = 1 - F_1(t)$. The observed event time $\tilde{T}_i$ is the time at which the first carcinogenic source becomes activated. In other words, $\tilde{T}_i = \min\{\zeta_k\}_{0\leq k\leq Y_i}$ for the ith subject. The population survival function $S_p(t)$ is defined as the probability of cancer non-detection at time t,

which is expressed as

$$S_p(t) = P(Y = 0) + P(\zeta_1 > t, \cdots, \zeta_{Y_i} > t; Y_i \geq 1) = exp\{-\theta(\mathbf{x}_i)F_1(t)\}. \qquad (2.6)$$

The population hazard function corresponding to (2.6) is $\lambda_p(t) = \theta(\mathbf{x}_i)f_1(t)$, where the density function is $f_1(t) = \frac{d}{dt}F_1(t)$. The cumulative hazard corresponding to (2.6) is defined as $\Lambda_p(t) = \int_0^t \theta(\mathbf{x}_i)f_1(z)dz = \theta(\mathbf{x}_i)F_1(t)$. I therefore rewrite Equation (2.6) as $S_p(t) = \exp\{-\Lambda_p(t)\}$. As $t \to \infty$, $S_p(t) \to \exp\{-\theta(\mathbf{x}_i)\} > 0$, where $S_p(t)$ is typically not a proper survival function.

Following Tsodikov (1998), I introduced a PH structure into Model (2.6): $S_p(t|\mathbf{x}_i) = \exp\{-F_1(t)\}^{\exp(\beta^T\mathbf{x}_i)} = S_{p,0}(t)^{\exp(\beta^T\mathbf{x}_i)}$. Suppose $S_{p,0}(t) = \exp\{-F_1(t)\}$, one could regard it as the baseline survival function associated with $F_1(t)$.

Variable Selection and Estimation

I first introduce an adaptive LASSO method for the promotion time cure model. Let the observed time $T_i = \min(\tilde{T}_i, C_i)$, where C_i is the non-informative and random censoring time. The censoring indicator $\delta_i = 1$ if $\tilde{T}_i \leq C_i$, and $\delta_i = 0$ otherwise. Model (2.6) has one set of independent variables $\mathbf{x}_i$ for subject i. For Model (2.6), the observed likelihood is

$$L(\beta, \alpha) = \prod_{i=1}^{n} \lambda_p(t_i)^{\delta_i} S_p(t_i) = \prod_{i=1}^{n} \left\{ \{\exp(\beta^T\mathbf{x}_i)f_1(t_i|\alpha)\}^{\delta_i} \exp\{-F_1(t|\alpha)\}^{\exp(\beta^T\mathbf{x}_i)} \right\} \qquad (2.7)$$

where α is the parameter in $F_1(\cdot)$.

For variable selection and parameter (β) estimation, I develop an EM-algorithm based on $(t_i, \delta_i, \mathbf{x}_i, y_i)$, where y_i is value of the Poisson cell count,Y_i. The log-likelihood

function for the complete data is

$$l_{pc}(\beta,\alpha;y) = \sum_{i=1}^{n}\Big\{\delta_i log(y_i f_1(t_i|\alpha)) + (y_i-\delta_i)log(1-F_1(t_i|\alpha)) + y_i\beta^T\mathbf{x}_i - e^{\beta^T\mathbf{x}_i} - logy_i!\Big\}. \tag{2.8}$$

For variable selection, I use an adaptive LASSO with the following penalized log-likelihood function:

$$pl_{pc}(\beta,\alpha;y) = \Big\{l_{pc}(.) - \tau^*\sum_{j=1}^{p}\frac{|\beta_j|}{|\rho_j|}\Big\}. \tag{2.9}$$

As in the case of mixture models, the tuning parameter τ^* determines the amount of penalty in Equation (2.9) and ρ functions as weights. Similarly, I obtain a consistent estimate of β by maximizing (2.8), and use it as the weight. When $\rho = 1$, this penalized function reduces to the familiar LASSO penalized function.

Computation

Let $(\beta^{(m)},\alpha^{(m)})$ be the parameter estimates in the mth iteration. To maximize Equation (2.9) for given τ^*, the EM algorithm takes the following steps:

E step: In the $(m+1)$th iteration, I compute $y_i^{(m+1)} = exp(\beta^{(m)T}x_i)\Big(1-F_1(t_i|\alpha^{(m)})\Big)$, and replace y_i in (2.9) with $y_i^{(m+1)}$.

M step: Solve the penalized score equation $U_P(\beta)$ for $\beta^{(m+1)}$ of β by using quadratic approximation (Fan and Li, 2001):

$$0 \;=\; U_P(\beta) = \sum_{i}^{n}\Big[y_i^{(m+1)} - exp(\beta^T\mathbf{x}_i)\Big]\mathbf{x}_i^T - \tau^*\sum_{j=1}^{p}\frac{\beta/|\beta^{(m)}|}{|\rho_j|}.$$

The penalized Hessian matrix H_β^* for β at $(m+1)$th iteration is given by $H_\beta^* = \frac{\partial}{\partial\beta}U_P(\beta)$.

In the M-step, to estimate α, I partition the time interval into non-overlapping

sub-intervals defined by $0 < s_1 < s_2 < ... < s_G$, with $s_G > max\{t_i\}$. I assume that $F_1(t|\alpha)$ follows a piecewise exponential model for which the hazard $\alpha_g (g = 1, 2, \cdots, G)$ remains constant for each sub-interval (Chen and Ibrahim, 2001). It can be shown by maximizing (2.8) with respect to α_g that for $i = 1, 2, \cdots, n$

$$\alpha_g^{(m+1)} = \Big[\sum_{s_{g-1} < t_i \leq s_g} \delta_i \Big] \times \Big[\sum_{s_{g-1} < t_i \leq s_g} y_i^{(m+1)}(t_i - s_{g-1}) + \sum_{y_i > s_g} y_i^{(m+1)}(s_g - s_{g-1}) \Big]^{-1}.$$

Alternatively, I can use the empirical distribution of $F_1(t)$ by assigning a point mass at each distinct observed event time so that $\sum f_1(t) = 1$ over the entire range of t. Suppose I have D distinct event times defined by $t_1^* < \cdots < t_D^*$. Let $f_1(t_d^*) = \alpha_d$ for $d = 1, 2, \cdots D$ so that $F_1(t_i|\alpha) = \sum_{t_d^* \leq t_i} \alpha_d$. For given values of $\beta^{(m)}$, I maximize (2.7) as a function of α only. The function to be maximized is

$$L_{\beta^{(m)}}(\alpha_1, \cdots, \alpha_D) \propto \prod_{d=1}^{D} \alpha_d \times exp\{-\alpha_d \sum_{i \in R_d} exp(\beta^{(m)T} \mathbf{x}_i)\}$$

The profile ML estimate $\alpha^{(m+1)}$ of α is given by

$$\alpha_d^{(m+1)} = \frac{1}{\sum_{i \in R_d} exp(\beta^{(m)T} \mathbf{x}_i)},$$

where R_d is the number of individuals at risk at time t_d^*. This yields an estimate of $F_1(t|\alpha)$

$$F_1^{(m+1)}(t|\alpha) = \sum_{t_d^* \leq t} \alpha_d^{(m+1)}$$

which is similar to the nonparametric version of the Breslow estimator of the baseline cumulative hazard.

The final estimator is obtained by iterating between the E and M steps until convergence. The EM algorithm has the following key steps:

Step 1: Determine an appropriate value for the tuning parameter τ^*, and initialize $\beta^{(0)}$

Step 2: Execute the E-step and estimate α

Step 3: Update the estimates as $\beta^{(1)} = \beta^{(0)} - H^{-1}(\beta^{(0)})U(\beta^{(0)})$

Step 4: Repeat steps 2 and 3 until $|\beta^{(1)} - \beta^{(0)}| \to 0$

For the tuning parameter selection, I use the same equation (2.5) to derive the BIC criterion for τ^*. Given $\hat{\beta}$ and $\hat{\alpha}$ I obtain an estimate of the log likelihood $\hat{l_{pc}}(.)$ from the unpenalized likelihood function. Using the BIC formula (2.5), I select a value of τ^* that minimizes the BIC.

As in mixture cure models, I take a two-step approach for parameter estimation and inference, i.e., independent variable effects are estimated and tested in a model with the selected variables. A standard approach for variance estimate is to use the inverse of the negative Hessian matrix derived from the observed likelihood (2.7). The covariance estimators for $\hat{\beta}$ given $F_1(t_i|\hat{\alpha})$ is

$$H(\hat{\beta}) = -F_1(t_i|\hat{\alpha})e^{\hat{\beta}^T\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i^T(e^{\hat{\beta}^T\mathbf{x}_i})^T.$$

In this research, I use bootstrap standard deviations for inference.

2.3 Simulation Study

I conduct a simulation study to evaluate the selection performance of the two cure rate models. Specifically, I compare the rates of selection accuracy of the proposed LASSO and adaptive LASSO methods, against that of the naïve p-value selection method. The significance level for the p-value procedure is set at 0.05, i.e., a variable is retained if the corresponding p value is less than 0.05.

2.3.1 Mixture Cure Rate Models

Data generation. I consider a scenario where $\mathbf{x} = (x_1, \cdots, x_9)^T$ has 9 independent variables. Three of the nine, x_5, x_6, x_9, are independent binary variables (1 vs 0) with probability $P(x_j = 1) = 0.5$, $j = 5, 6, 9$. The other independent variables in $\mathbf{x}$ are standard normally distributed with a pairwise correlation between x_i and x_j of $\rho^{|i-j|} = 0.5$, which reflects a moderately strong correlation. For the logistic component of the mixture cure model, vectors of regression coefficients are set to $\beta = (0.5, 0.10, -0.25, 0, 0, 0, 0, 0, 0)^T$. For the survival model, I assume without loss of generality that $\mathbf{x} = \mathbf{z}$. Failure times are generated from a Weibull distribution with a survival function $S(t|a, b) = \exp\{-\frac{t}{b}\}^a$. The shape parameter is $a = 1.5$, and scale parameter $b = \exp\{e^{\gamma^T \mathbf{z}}\}^{-1/a}$ with $\gamma = (1, 0, 0.1, 0.25, 0, 0, 0, 0, 0)^T$. I include an intercept for the logistic model, and no intercept for the survival model. The mean cure rate is approximately 30%. Censoring times are generated from Uniform(c, d), where c and d are selected to achieve the desired censoring rate. I considered two different levels of censoring: 20% and 50%.

For each parameter setting, I generated 100 datasets, with sample sizes of 250 and 500. I apply the LASSO, adaptive LASSO, and naïve p-value procedures for variable selection. I implement the selection procedure with both parametric and nonparametric estimators for the baseline hazard $\Lambda_{nc,0}(t)$. I apply the penalized methods for variable selection with given values of the tuning parameter $\tau = (\tau_1, \tau_2)$ for the mixture cure rare model. Optimal values of the tuning parameter are selected by minimizing the BIC selector (2.5).

Simulation results. Table 2.1 presents the selection results for the mixture cure model. Six elements of β and γ have zero effects, whereas the other three have nonzero effects. I present the average number of correct exclusion (unimportant effects not being selected) and the average number of incorrect exclusion (important effects not

being selected) for the logistic regression coefficients β and PH regression coefficients γ. The table summarizes the results based on 100 simulations.

Briefly, for the logistic regression component in the mixture model, the rate of incorrect exclusion is zero for both LASSO and adaptive LASSO. In other words, both regularization methods have correctly included all three non-zero effects. In comparison, the p-value method has on average incorrectly excluded $2.96 - 3$ of the 3 nonzero effects, a very poor performance by any standard. In the meantime, the adaptive LASSO has excellent rates of correct exclusion: On average, it is able to exclude $4.8 - 6$ of the 6 true zero effects. This performance is similar to that of the p-value method which consistently excludes $5.8 - 6$ of the zero effects. The LASSO method, on the other hand, tends to exclude fewer zero effects.

For the PH component, all three methods have correctly included the three non-zero effects. The difference is in the exclusion of zero effects. In this regard, The adaptive LASSO has the best performance. It is able to exclude $3.66 - 5.91$ of the 6 zero effects. LASSO has slightly worse but still acceptable performance. The p-value method, on the other hand, completely fails to exclude any of the zero effects.

In comparing the selection performance of the the balancing the two different types of errors, the adaptive LASSO appears to outperform its competitors. Importantly, the superior performance of the adaptive LASSO procedure is consistent across all simulation settings and it does not appear to be greatly influenced by the censoring proportion and how baseline hazards are estimated.

2.3.2 Promotion Time Cure Rate Model

Data generation. For the promotion time cure model, I again consider a situation where $\mathbf{x} = (x_1, \cdots, x_9)^T$ has nine independent variables. Three of the nine, x_5, x_6, x_9, are independent Bernoulli variables with probability $P(x_j = 1) = 0.5$, $j = 5, 6, 9$. The other six variables of $\mathbf{x}$ are standard normally distributed with a pairwise correlation

between x_i and x_j of $\rho^{|i-j|} = 0.5$. As in Model (2.6), I assume that the mean number of cancer cells is $\theta = \exp(\beta^T \mathbf{x})$ with $\beta = (0.5, 0.10, -0.25, 0, 0, 0, 0, 0, 0)^T$. I also assume that $F_1(t)$follows a Weibull distribution with scale parameter $b = \exp\{\theta\}^{-1/a}$, and the shape parameter $a = 1.5$. Censoring times are generated from a uniform distribution yielding censoring rate of 20% and 50%. I generated 100 datasets for each setting with sample sizes 250 and 500, and censoring percentages 20% and 50%. I fit Model (2.6) by using both parametric and nonparametric estimates of $F_1(t)$.

Simulation results. Table 2.2 depicts the selection results for promotion time cure model. The simulation shows that the adaptive LASSO outperforms both LASSO and the p-value methods in identifying the zero effects, as evidenced by its high correct exclusion rates, while maintaining a perfect rate of including all non-zero effects. The LASSO has respectable performance in achieving a perfect rate of including all non-zero effects, but it is slightly less effective in identifying the zero effects. The p-value method tends to incorrectly exclude the true non-zero effects at the unacceptable rates of $1.28 - 1.98$ out of 3.

2.3.3 Post-Selection Inference

In the absence of formal theoretical development of post-selection inference, analysts are likely to perform inference based on the selected model. Here I conduct a simulation study to examine the empirical performance of the practice. Specifically I examine the 95% coverage probabilities and the average bootstrap standard errors (ASE) for the nonzero coefficients of β and γ. Here bootstrap standard errors are obtained based on 100 resamples. Simulation results are presented in Table 2.3. Briefly, the coverage probabilities are generally good, especially for the promotion time cure rate models, even with 50% censoring. The performance of the mixture model is slightly more variable. Overall, the simulation seems to provide some empirical evidence in support of the two step selection-estimation procedure.

Finally, I conducted a sensitivity analysis examining the selection performance in misspecified models, i.e., data are generated from mixture models when promotion time models are fitted, or vice versa. In the strictest sense, the mixture cure rate model (MCM) and promotion cure rate model (PCM) are not directly comparable because of their differences in structure and assumption. In the case A (MCM as true model) and Case B (MCM as misspecified model), Table 2.4 only shows the survival component of the MCM fits. Simulation shows that selection accuracies in the survival components of the true and mis-specified models were generally comparable, which provided some assurance on the robustness of the selection method. But I caution against over-interpretation because the simulation has not taken into account the selection performance of the logistic component in the MCM. Detailed results are included in Table 2.4.

2.4 Application

To illustrate the proposed methods, I consider a real clinical investigation of childhood wheezing. The basic study design was described elsewhere (Tepper *et al.*, 2008). Briefly, this is an observational study aimed at understanding the risk factors associated with early onset of wheezing. For this purpose, the variable selection methods that I develop provided a logical tool for risk factor screening. The onset age of wheezing symptoms was the main outcome of interest. Onset age was determined from the monthly reports of wheezing episodes during the study period. The study recruited a total of 116 children. Enrolled children were followed prospectively for up to 5 years. Eighty-six ($n = 86$) children completed the designed follow-up. The current analysis was based on data from these 86 children with complete follow-up.

A total of 13 variables were considered in the current analysis. The demographic and general health variables included race (RACE, 1=white and 0=non-white), sex (GENDER, 1=male, 0=female), and mother's smoking status during

pregnancy (1=nonsmoker mother during pregnancy, and 0=otherwise), allergy to food (FOODANT; 1=yes, 0=no), egg or milk (EGGMILK, 1=yes, 0=no), and use of topical steroids (TOPSTEO, 1=yes, 0=no). Continuous variables included: (1) provocative concentration of methacholine corresponding to 30% drop in forced expiratory volume in 1 second (logPC30 (mg/ml)); (2) centralized height (CenHEIGHT (cm)); (3) severity of eczema, a score ranged from 0 to 29 calculated based the levels of body surface involvement, intensity of symptom, and presence of pruritus and insomnia (SCVALUE); (4) logarithmic transformed level of total serum immunoglobulin E (log(ITOTAL)); (5) Z-score of forced vital capacity (ZFVC), (6) Z-score of forced expiratory flow $25\% - 75\%$ (ZFEF2575); (7) Z-score of forced expiratory volume in half a second (ZFEV5). Among these, the last three variables (ZFVC, ZFEF2575, and ZFEV5) were lung function measurements. The average age at enrollment of these children was approximately 10.7 months. The median age at the first wheeze episode was 21.67 months. Summary statistics of the independent variables are reported in Table 2.5.

Kaplan-Meier estimates of the wheezing free probabilities for boys and girls are presented in Figure 2.1. The Kaplan-Meier plot for girls flattened after 48 months, with relatively few censoring, suggesting that a portion of the population were not subject to any risk of wheezing. A similar pattern was seen in boys. To accommodate this fraction of the cured, I analyzed the data using a mixture cure rate model (2.1). I did not consider promotion time cure models in the absence of a clear biological rationale for that approach. Wheezing, as an airway symptom, does not have a single and specific cause that justifies the use of a promotion time model. I performed variable selection using methods described in the paper. Both LASSO and adaptive LASSO methods were used.

To select the tuning parameters for the logistic regression and PH regression models, for a given set of tuning parameter values I plug in the estimates $\hat{\beta}$ and $\hat{\gamma}$ into

Equation (2.3). And then I optimize the tuning parameters that minimize the BIC selector (2.5). Under the LASSO penalty, all 13 variables were retained for the logistic regression model. The adaptive LASSO produced a more parsimonious logistic model with 5 independent variables: SCVALUE, GENDER, RACE, MONSMOKE, and TOPSTEO. For the PH model, the LASSO penalty selected 11 of the 13 variables: GENDER, RACE, MOMSMOKE, FOODANT, EGGMI LK, TOPSTEO, ZFEF2575, ZFEV5, HEIGHT, SCVALUE, and ITOTAL. The adaptive LASSO selected 7 variables: SCVALUE, GENDER, RACE, MONSMOKE, FOODANT, EGGMILK, and TOPSTEO. I present the final model fitting results based on the adaptive LASSO method in Table 2.6. Of note, the model identified by the adaptive LASSO was more parsimonious, and it included all of the variables identified by the LASSO method.

A careful examination of the parameter estimates from the selected model revealed that: (1) an estimated $49\% = 1/(1 + 1.03)$ of population were subject to the risk of wheezing if all other factors (SCVALUE, GENDER, RACE, MOMSMOKE, and TOPSTEO) were set to 0; (2) male sex, white race, mother smoked during pregnancy, topical steroid use, and greater eczema severity were associated with increased risk of wheezing. For the children who were at risk, a greater eczema severity, mother smoking during pregnancy, white race, male sex, topical steroid use, and known allergy to food, egg, and milk were associated with early onset of wheezing.

2.5 Discussion

Cure rate model represents an important class of methods for analyzing time-to-event data, in situations where certain individuals are free of the disease risk. Because of the increased complexity in modeling structure, a common challenge that analysts face is the determination of model composition, i.e., what independent variables should be included in or excluded from which modeling components. While fully subjective variable selection by investigators is usually thought to be error-prone, the traditional

p value-based selection methods are not always efficient and stable. To alleviate the challenge, I present two selection methods, based on LASSO and adaptive LASSO, to aid variable selection in different types of cure rate models. Built on earlier attempts on the mixture cure model (Liu *et al.*, 2012), this work further extends the selection tool to promotion time models. Extensive simulation shows that the adaptive LASSO method has superior performance than the LASSO and p-value methods, in terms of selection accuracy. The method appears to have worked well for both mixture and promotion time cure rate models. Making these methods available to practitioners, I hope, would have an impact on how cure rate models are used in analytical practice. The selection of independent variables are of course not limited to main effects, two-way or higher order interactions can be incorporated with modification of the design matrices for the logistic and survival components, with the usual understanding that the main effects are to be included if an interaction involving them is selected. Computationally, as I have demonstrated in the current paper, adaptive LASSO is generally efficient, and it is easily implementable in various computing platforms.

A few practical issues deserve some discussion: (1) Determination of the initial sets of independent variables going into the logistic and survival components is generally guided by subject science, and it typically reflects the investigators' understanding of the cure and survival processes. In the absence of strong scientific reasons for including and/or excluding certain variables into the initial sets of independent variables, analysts typically use the same set of variables for both components, so $\mathbf{x} = \mathbf{z}$ is a rather common practice.(2) Estimation of the unknown baseline hazard functions. Previously, different authors have explored various approaches. Among the published methods, for mixture cure rate model Sy and Taylor (2000) used a Breslow type estimator and a product limit estimator, Farewell (1982) considered a parametric (Weibull) model, Corbiére *et al.* (2009) attempted the use of nonparametric spline functions, and Chen and Ibrahim (2001) used a piecewise exponential model

for hazard function for promotion time cure model. In this research, I constructed a nonparametric step-function for baseline hazard under the promotion time cure model. For mixture cure rate model I utilized a piecewise constant hazard function for baseline hazard approximation. I compared the performance of variable selection of adaptive LASSO and LASSO using the Breslow type estimator and piecewise exponential model for the baseline hazard function in the simulation. My simulation shows that different choices of baseline hazard estimators produced generally comparable selection results. Considering the simplicity of the approach concludes that the choice of baseline hazard estimation methods is not as consequential as previously thought, at least for the purpose of variable selection. (3) Determination of the weights for adaptive LASSO. Ideally, the weights need to be data-dependent and consistent with the oracle properties (Fan and Li, 2001). When the number of variables is larger and many of them are correlated, the consistent estimates may be difficult to obtained. Thus the issue requires further investigation. (4) Estimation of standard errors. Standard error estimates are important for the purpose of inference. For linear models Tibshirani (1996) and Fan and Li (2001) provided Hessian matrix-based standard error estimates, while Zou (2006) advocated the use of bootstrap estimates. For nonlinear models, penalized variable selection methods tend to introduce biases in the estimation of model parameters. The magnitude of the bias is influenced by the choice of weights or tuning parameters. As a result, Hessian matrix-based standard error estimates do not work well for inference, at least in my modeling setting (data not shown). So in this research, I chose to use bootstrap standard error estimates in the selected model, to minimize the impact of tuning parameters and thus alleviating the risk of estimation bias. (5) Post-selection inference. As stated earlier, this paper has primarily focused on variable selection and not on post-selection inference. The two-stage estimation process is somewhat an *ad hoc* way to obtain the approximation of standard errors, but it is generally consistent with the current

biostatistical practice of making inference based on the final selected models (Moore and McCabe, 2009). Most recently, Berk and colleagues (2013) have suggested that one could reframe the post-selection testing in the context of simultaneous inference, which takes into account the multiplicity associated with all sub-models (all linear functions of estimates) instead of the selected model, in hoping that the inference no longer depends on correct selection of the true model. Berk's approach was discussed in a linear model setting. Extension of this approach to nonlinear settings remains to be developed. In the absence of rigorous methodological development, I opted for the standard two-step approach. The simulation study seems to support the notion of a generally good selection performance, at least in tested settings. On balance, use of resampling in a two-step process, in opinion, represents a sensible compromise between accurate standard error estimation and valid inference performance. It has been shown in a previous work that such a method works well in complex modeling settings (He *et al.*, 2015).

Tables:

Table 2.1: Simulation study. Performance of variable selection results for mixture cure model with 20% and 50% censoring. The average numbers of correct exclusion (exclusion of zero effects) and incorrect exclusion (exclusion of non-zero effects)

	20% censoring	Average number of 0 coefficients			
n	Method	β(logistic)		γ(survival)	
	Nonpar $\lambda_0(t)$	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
250	Oracle	6	0	6	0
	$P < 0.05$	6	2.96	0	0
	Adaptive LASSO	5.23	0	4.68	0
	LASSO	4.97	0	4	0
500	Oracle	6	0	6	0
	$P < 0.05$	6	2.98	0	0
	Adaptive LASSO	6	0	5.90	0
	LASSO	6	0	4.57	0
	Par $\lambda_0(t)$				
250	Oracle	6	0	6	0
	$P < 0.05$	5.98	3	0	0
	Adaptive LASSO	5.76	0	4.34	0
	LASSO	4	0	3.86	0
500	Oracle	6	0	6	0
	$P < 0.05$	5.99	3	0	0
	Adaptive LASSO	6	0	5.91	0
	LASSO	4.52	0	4	0
	50% censoring	β(logistic)		γ(survival)	
	Nonpar $\lambda_0(t)$	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
250	Oracle	6	0	6	0
	$P < 0.05$	5.80	2.90	0	0
	Adaptive LASSO	5.02	0	4.02	0
	LASSO	3.99	0	3.76	0
500	Oracle	6	0	6	0
	$P < 0.05$	4.92	2.98	0	0
	Adaptive LASSO	6	0	4.99	0
	LASSO	6	0	3.71	0
	Par $\lambda_0(t)$				
250	Oracle	6	0	6	0
	$P < 0.05$	5.88	2.94	0	0
	Adaptive LASSO	4.80	0	3.66	0
	LASSO	3.75	0	3.66	0
500	Oracle	6	0	6	0
	$P < 0.05$	6	3	0	0
	Adaptive LASSO	6	0	5.69	0
	LASSO	4.41	0	4	0

Table 2.2: Simulation study. Performance of variable selection results for promotion time cure model with 20% and 50% censoring. The average numbers of correct exclusion (exclusion of zero effects) and incorrect exclusion (exclusion of non-zero effects)

n	Method	Average number of 0 coefficients			
		20% censoring		**50% censoring**	
		β		β	
	Nonparametric specification	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
250	Oracle	6	0	6	0
	$P < 0.05$	4.42	1.46	5.55	1.83
	Adaptive LASSO	6	0	6	0
	LASSO	4.62	0	4.30	0
500	Oracle	6	0	6	0
	$P < 0.05$	4.85	1.56	5.92	1.98
	Adaptive LASSO	6	0	6	0
	LASSO	5.64	0	5.33	0
	Parametric specification	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
250	Oracle	6	0	6	0
	$P < 0.05$	4.72	1.48	4.65	1.48
	Adaptive LASSO	6	0	4.45	0
	LASSO	3.98	0	3.45	0
500	Oracle	6	0	6	0
	$P < 0.05$	4.44	1.28	5.48	1.89
	Adaptive LASSO	6	0	6	0
	LASSO	4	0	3.83	0

Table 2.3: Simulation study. Empirical 95% coverage probability (Coverage prob), and average values of the estimated bootstrap standard errors (ASE) of the estimates in the Adaptive LASSO selected models. MCM stands for mixture cure rate models, and PCM stands for promotion time cure rate model.

n	Model	Coefficient	20% censoring		50% censoring	
			Coverage prob.	ASE	Coverage prob.	ASE
250	MCM–logistic	β_1	0.82	0.090	0.91	0.203
		β_2	0.95	0.072	0.96	0.148
		β_3	0.95	0.036	0.96	0.074
250	MCM–survival	γ_1	0.96	0.024	0.95	0.022
		γ_2	0.96	0.029	0.91	0.024
		γ_3	0.96	0.014	0.91	0.011
500	MCM–logistic	β_1	0.88	0.080	0.89	0.071
		β_2	0.95	0.057	0.96	0.068
		β_3	0.95	0.028	0.96	0.034
500	MCM–survival	γ_1	0.96	0.021	0.98	0.139
		γ_2	0.94	0.022	0.98	0.150
		γ_3	0.94	0.011	0.98	0.075
250	PCM	β_1	0.93	0.002	0.95	0.010
		β_2	0.95	0.004	0.95	0.010
		β_3	0.95	0.002	0.95	0.005
500	PCM	β_1	0.95	0.002	0.96	0.003
		β_2	0.95	0.003	0.96	0.003
		β_3	0.95	0.002	0.96	0.002

Table 2.4: Sensitivity analysis on robustness of model misspecification. PCM stands for promotion time cure model, and MCM stands for mixture cure rate model.

		CASE A(true model MCM and fitted as PCM)			
		20% censoring		**50% censoring**	
n	Method	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
250	Oracle	6	0	6	0
	PCM-Adp. Lasso	5.26	0	5.39	0
	PCM-Lasso	4.58	0	4.19	0
	MCM-Adp. Lasso	4.68	0	4.02	0
	MCM-Lasso	4	0	3.76	0
500	Oracle	6	0	6	0
	PCM-Adp. Lasso	5.83	0	5.90	0
	PCM-Lasso	5.47	0	5.05	0
	MCM-Adp. Lasso	5.90	0	4.99	0
	MCM-Lasso	4.57	0	3.71	0

		CASE B(true model PCM and fitted as MCM)			
		20% censoring		**50% censoring**	
n	Method	Correct exclusion (6)	Incorrect exclusion (3)	Correct exclusion (6)	Incorrect exclusion (3)
250	Oracle	6	0	6	0
	PCM-Adp. Lasso	6	0	6	0
	PCM-Lasso	4.62	0	4.30	0
	MCM-Adp. Lasso	4	0	4	0
	MCM-Lasso	4	0	4	0
500	Oracle	6	0	6	0
	PCM-Adp. Lasso	6	0	6	0
	PCM-Lasso	5.64	0	5.33	0
	MCM-Adp. Lasso	4.91	0	4.08	0
	MCM-Lasso	4.98	0	4	0

Table 2.5: Baseline characteristics of subjects included in the analysis

Factor	n	Variable	Mean	Variance
$GENDER$	0 42	$ZFVC$	-0.319	1.222
	1 44			
$RACE$	0 43	$ZFEF2575$	-0.689	0.837
	1 43			
$MOMSMOKE$	0 9	$ZFEV5$	-0.614	1.108
	1 77			
$FOODANT$	0 55	$CenHEIGHT$	-0.673	41.935
	1 31			
$EGGMILK$	0 59	$SCVALUE$	9.547	50.203
	1 27			
$TOPSTEO$	0 47	$log(ITOTAL)$	2.147	2.700
	1 39			
		$logPC30$	-0.787	1

Table 2.6: Summary of parameter estimates with confidence intervals and two sided p-values for the childhood wheezing study. In the logistic model, OR stands for odds ratio. In the survival model, HR refers to hazard ratio.

Variable	OR (CI)	p-value	HR (CI)	p-value
Intercept	1.030 (1.017, 1.042)	0.000		
$SCVALUE$	1.337 (1.242, 1.439)	0.000	1.275 (1.196, 1.400)	0.000
$MOMSMOKE$	1.025 (1.014, 1.037)	0.000	1.022 (1.011, 1.034)	0.000
$RACE$	1.016 (1.010, 1.025)	0.000	1.014 (1.007, 1.022)	0.000
$GENDER$	1.017 (1.010, 1.025)	0.000	1.016 (1.010, 1.024)	0.000
$TOPSTEO$	1.013 (1.010, 1.018)	0.000	1.010 (1.005, 1.015)	0.000
$FOODANT$			1.009 (1.003, 1.016)	0.002
$EGGMILK$			1.008 (1.002, 1.014)	0.005

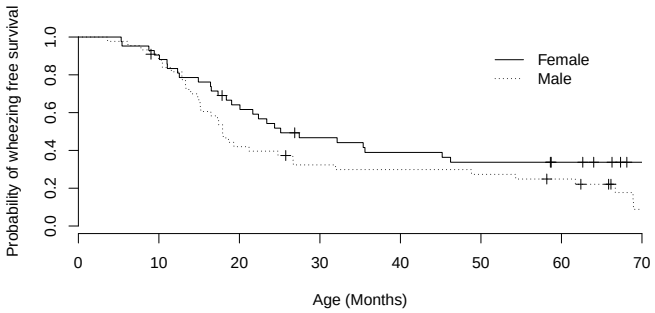


Figure 2.1: Kaplan-Meier estimates of wheezing-free probabilities in male and female subjects

Chapter 3

OPTIMAL MODEL SELECTION FOR PARTIALLY LINEAR MIXTURE CURE RATE MODELS

This chapter presents a penalized based selection procedure for a partially linear mixture cure rate model by extending the mixture cure rate model proposed in Chapter 2. The partially linear modeling framework, which is is a flexible setup, accommodates linear and nonlinear variables within one model structure. A central key question about understanding the underlying hidden functional structure in independent variables is an essential element for modeling any complex data with the partially linear models. The selection procedure is developed and applied it to a childhood hypertension study data. A relevant inference procedure is examined by various simulation studies and illustrated by the real clinical data. An expected-maximization (EM) algorithm is developed to implement the method. Thus, the proposed model selection technique expresses a great demand of application on a generalized modeling setting.

3.1 Research Background

In survival analysis, when a subset of subjects are not at risk of the event of interest, analysts typically introduce a logistic component to accommodate the proportion of subjects that are not at risk, i.e., those who are "cured" of the underlying disease. Such models are often referred to as the cure rate models, which simultaneously accommodate the cured and the non-cured subjects. The model expresses the probability of being cured as $1-\theta(\boldsymbol{x})$, and the survival function of the non-cured as $S_{nc}(t|\boldsymbol{z})$ (i.e., $\lim_{t\to\infty} S_{nc}(t|\boldsymbol{z}) = 0$), where $\theta(\boldsymbol{x})$ is the failure probability (Boag, 1979; Farewell,

1977, 1982). The population survival function, $S_p(t)$, can therefore be expressed as

$$S_p(t) = 1 - \theta(\boldsymbol{x}) + \theta(\boldsymbol{x}) S_{nc}(t|\boldsymbol{z}), \tag{3.1}$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are respectively vectors of independent variables that influence the cure rate, and the survival of the non-cured. In Model (3.1), $S_{nc}(t|\boldsymbol{z})$ describes the time-to-event distribution of the non-cured subjects, and $1-\theta(\boldsymbol{x})$ quantifies the cure probability. In practice, $\theta(\boldsymbol{x})$ is often modeled by a logistic model and $S_{nc}(t|\boldsymbol{z})$ by a Cox proportional hazard model (Cox, 1972) . Traditional cure rate models assume that both $\boldsymbol{x}$ and $\boldsymbol{z}$ are linear effects, although the validity of the assumption is rarely verified.

The basic structure of (3.1) has later been extended to accommodate nonlinear effects to minimize the risk of model misspecification (Hastie and Tibshirani, 1990; Hastie *et al.*, 2001). The extended models are referred to as the additive cure rate models (Wang *et al.*, 2012; Corbiére *et al.*, 2009) . When a model contains additive nonlinear effects, the task of variable selection must be expanded to determine (1) whether a variable should be included, and (2) if so, whether it should be included as a linear or nonlinear effect. This expanded task is the topic of the current research. I propose a selection procedure that simultaneously selected variables from $\boldsymbol{x}$ and $\boldsymbol{z}$.

Penalized likelihood methods have been used for variable selection in various modeling situations, following Tibshirani's work on the Least Absolute Shrinkage and Selection Operators (LASSO) (Tibshirani 1996, 1997). Several authors have investigated the oracle property of the approach in various modeling situations, including linear models (Fan and Li, 2001; Zou, 2006), Cox models (Zhang and Lu, 2007), and joint models of linear and survival outcomes (He *et al.*, 2015). Zou (2006) proposed an adaptive LASSO penalty that uses data-driven weights for penalizing different coefficients, and thus shrinking small coefficients to zero. In this research, I use an

adaptive LASSO penalty to determine the model composition in mixture cure rate models.

How nonlinear effects should be selected, to some extent, depends on the way that these effects are modeled. Literature abounds when it comes to nonlinear modeling techniques. Among the available methods, various spline-based regression are increasingly been used in analytical practice (O'sullivan, 1986; Wahba, 1990; Gray, 1992; Eilers and Marx,1996; Wang, 2011). For practical purposes, the different smoothing techniques often makes little difference in the fitness of the resultant model. Herein, I use B-spline techniques for their numerical stability (Eilers and Marx, 1996, P̃ena, 1997). When it comes to detection of nonlinear effects, Zhang and colleagues (2011) proposed a data-driven method in a linear model setting. For Cox regression models, Liu and colleagues (2012) further extended the variable selection methods to a cure rate model setting. But the non-convex of their penalty tended to lead to numerical instability (Zhang and Lu, 2007). To remedy, I propose to use the adaptive LASSO penalty, as described by Zou (2006) and Zhang and Lu (2007), that is computationally stable for the purpose of variable selection in cure rate models.

Herein, I consider an additive mixture cure rate model with unspecified functional forms of independent variables as depicted by cubic B-splines. By partitioning the nonparametric functions into linear component and nonlinear component, I use an L_1 (or, LASSO) penalty in model selection. This approach is linked to the work by Wand and Ormerod (2008) who decomposed fixed and random effects into a mixed effect model. I selected linear effects as fixed effects, and nonlinear effects as random effects. To implement, I use an expectation-maximization (EM) algorithm. Extensive simulation studies are conducted to evaluate operational characteristics of the proposed method. Finally, I illustrate the use of the method by analyzing data from a real clinical study.

3.2 Mixture Cure Rate Models

3.2.1 Model Formulation

Let $\tilde{T_i}$ and C_i the underlying failure time and censoring time respectively for ith subject, $i = 1, 2, ..., n$. Observed time is $T_i = \min(\tilde{T_i}, C_i)$. I assume C_i is random, noninformative, and is independent of $\tilde{T_i}$. Let the failure indicator $\delta_i = 1$ if $\tilde{T_i} \leq C_i$ ($\tilde{T_i}$ is observed), and $\delta_i = 0$ (censored) otherwise .

I denote the independent variables in the logistic model as $\boldsymbol{x_i} \in \mathbb{R}^p$, and the variables in the survival model as $\boldsymbol{z_i} \in \mathbb{R}^q$, where $p \geq q$. In most applications, $\boldsymbol{x_i}$ and $\mathbf{z_i}$ have common elements. Let Y_i be a binary indicator for non-cured, and the probability is $P(Y_i = 1) = \theta(\boldsymbol{x_i})$; therefore $\{1 - \theta(\boldsymbol{x_i})\}$ is the proportion of cured subjects. Under such a notation, the population survival function $S_p(t)$ is given by $S_p(t) = \{1 - \theta(\boldsymbol{x_i})\} + \theta(\boldsymbol{x_i}) S_{nc}(t|\boldsymbol{z_i})$, where $S_{nc}(t)$ is survival function of the non-cured, given $\boldsymbol{z_i}$. As t increases, $S_p(t) \to \{1 - \theta(\boldsymbol{x_i})\} > 0$. Note that $S_p(t)$ is not a proper survival function.

Farewell (1982) described the effects of independent variables $\boldsymbol{x}$ on the probability of not being cured with a logit link function as

$$\log\left\{\frac{\theta(\boldsymbol{x_i})}{1 - \theta(\boldsymbol{x_i})}\right\} = \mathcal{A} + \sum_{h=1}^{p} f_h(\boldsymbol{x_{i,h}}),$$

where $\mathcal{A}$ is an unknown intercept term, and $f^T(x_i) = (f_1(x_{i,1}), f_2(x_{i,2}), \ldots, f_p(x_{i,p}))$ is a vector of nonparametric function of $\boldsymbol{x_i}$. I centralize the $\boldsymbol{x_i}$ to ensure identifiability. I model $f_h(.)$ with cubic-B spline functions.

Similarly for the non-cured, the Cox proportional hazard model are written as

$$\lambda_{nc}(t|\boldsymbol{z_i}) = \lambda_{nc,0}(t) \exp\{\sum_{m=1}^{q} g_m(\boldsymbol{z_{i,m}})\}.$$

I write $log\{\mu(\boldsymbol{z_i})\} = \sum_{m=1}^{q} g_m(\boldsymbol{z_{i,m}})$, where $\mu(\boldsymbol{z_i})$ is a nonnegative link function, and the vector of nonparametric function of z_i is $g^T(z_i) = (g_1(z_{i,1}), g_2(z_{i,2}), \ldots, g_q(z_{i,q}))$. Centralizing the $\boldsymbol{z_i}$ to ensure the identifiability, I again assume that $g_m(.)$ can be depicted by cubic-B spline functions.

In this formulation, the baseline hazard function is $\lambda_{nc,0}(t)$. The cumulative baseline hazard function is $\Lambda_{nc,0}(t) = \int_0^t \lambda_{nc,0}(u)du$, and the baseline survival function is $S_{nc,0}(t) = exp\{-\Lambda_{nc,0}(t)\}$. $\mu(\mathbf{z_i})$.

Plugging in the spline functions, I rewrite the additive model (3.1) as

$$S_p(t) = \{1 - \frac{e^{\{\mathcal{A}+\sum_{h=1}^{p} \boldsymbol{f}_h(\boldsymbol{x}_{i,h})\}}}{1 + e^{\{\mathcal{A}+\sum_{h=1}^{p} \boldsymbol{f}_h(\boldsymbol{x}_{i,h})\}}}\} + \{\frac{e^{\{\mathcal{A}+\sum_{h=1}^{p} \boldsymbol{f}_h(\boldsymbol{x}_{i,h})\}}}{1 + e^{\{\mathcal{A}+\sum_{h=1}^{p} \mathbf{f}_h(\boldsymbol{x}_{i,h})\}}}\} S_{nc,0}(t)^{\exp\{e^{\sum_{m=1}^{q} \boldsymbol{g}_m(\boldsymbol{z}_{i,m})}\}}, \tag{3.2}$$

3.2.2 Method

Denoting observed data as $(t_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, I write the likelihood function as

$$L(\boldsymbol{f(.)}, \boldsymbol{g(.)}, \lambda_{nc,0}) = \prod_{i=1}^{n} \{\theta(\boldsymbol{x}_i)\lambda_{nc,0}(t_i)\mu(\boldsymbol{z}_i)S_{nc,0}(t_i)^{\mu(\boldsymbol{z}_i)}\}^{\delta_i} \prod_{i=1}^{n} \{1 - \theta(\boldsymbol{x}_i)(1 - S_{nc,0}(t_i)^{\mu(\boldsymbol{z}_i)})\}^{1-\delta_i},$$

When the event time is censored, the cure information y_i is not observable. However, I can construct the complete likelihood based on $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i, y_i)$ and use an EM algorithm to maximize the complete likelihood function by treating y_i as a latent binary variable. The complete likelihood can be factored into a logistic model for the cured and a PH model for the non-cured:

$$L_C(\mathbf{f}(.), \boldsymbol{g(.)}, \lambda_{nc,0}; y) = \prod_{i=1}^{n} \theta(\boldsymbol{x}_i)^{y_i}(1 - \theta(\boldsymbol{x}_i))^{1-y_i} \prod_{i=1}^{n} \left[\{\lambda_{nc,0}(t_i)\mu(\boldsymbol{z}_i)\}^{\delta_i} S_{nc,0}(t_i)^{\mu(\boldsymbol{z}_i)}\right]^{y_i}.$$

The log likelihood function is

$$
\begin{aligned}
l_C(\boldsymbol{f(.)}, \boldsymbol{g(.)}, \lambda_{uc,0}; y) \;=\; & \sum_{i=1}^{n} \Big\{ y_i \{ \sum_{h=1}^{p} \boldsymbol{f}_h(\boldsymbol{x}_{i,h}) \} - log\{1 + e^{\sum_{h=1}^{p} \boldsymbol{f}_h(\boldsymbol{x}_{i,h})} \} \Big\} \\
& + \sum_{i=1}^{n} \Big\{ y_i \delta_i \{ log\{\lambda_{nc,0}(t_i)\} \\
& + \sum_{m=1}^{q} \boldsymbol{g}_m(\boldsymbol{z}_{i,m}) \} + y_i \exp(\sum_{m=1}^{q} \boldsymbol{g}_m(\boldsymbol{z}_{i,m})) \log\{S_{nc,0}(t_i)\} \Big\}.
\end{aligned} \tag{3.3}
$$

For simplicity, I write the first term and the rest of Equation (3.3) as $l_{C,1}(\mathbf{f}(.); \mathbf{y})$ and $l_{C,2}(\mathbf{g}(.), \lambda_{nc,0}; y)$. I therefore have $l_C(\boldsymbol{f(.)}, \boldsymbol{g(.)}, \lambda_{uc,0}; y) = l_{C,1}(\boldsymbol{f(.)}, \boldsymbol{y}) + l_{C,2}(\boldsymbol{g(.)}, \lambda_{nc,0}, y)$. In other words, the log likelihood of the cure rate model is the sum of the log likelihood of the logistic model and the log likelihood of the proportional hazard model.

3.2.3 Variable Selection

I approximate $\boldsymbol{f(x)}$ and $\boldsymbol{g(z)}$ with a cubic B-spline with K inner knots. The B-spline basis functions have many attractive features, including smoothness of the curve and numerical stability of the computation (Eilers and Marx, 1996; P̃ena, 1997). The approximations are $\boldsymbol{f(x_i)} = \boldsymbol{B(x_i)}\tilde{\boldsymbol{\beta}}$, and $\boldsymbol{g(z_i)} = \boldsymbol{B(z_i)}\tilde{\boldsymbol{\gamma}}$, where $\boldsymbol{B}$ is the $n \times (K+3)$ design matrix; $\tilde{\boldsymbol{\beta}}^T$ and $\tilde{\boldsymbol{\gamma}}^T$ are the coefficient vector of $\boldsymbol{B}(\boldsymbol{x}_i)$ and $\boldsymbol{B}(\boldsymbol{z}_i)$, respectively.

I use a penalized method to select independent variables and to discover potential nonlinear effects. The L_1 penalty is used in that regard. The L_1 penalty, however helpful in selecting variables, is not particularly useful for discriminating nonlinear effects, i.e., for structural discovery. To remedy, I decompose $f_h(.)$ and $g_m(.)$ into the linear and nonlinear components and then use the adaptive LASSO selection approach to identify the effects. It is an approach that simultaneously selects important variables and identifies the true structure of the variables by giving different weights on different coefficients in the L_1 penalty.

Wand and Ormerod (2008) used a similar spectral decomposition technique to obtain an exact matrix expression using the O'Sullivan spline (O'sullivan, 1986, 1988; Hastie and Tibshirani, 1990). O'Sullivan spline is a class of B-spline functions. Cubic B-spline falls into such a category. Briefly, Wand and Ormerod (2008) assume that $\boldsymbol{U}$ is the eigenvector of a $(K+3)\times(K+3)$ matrix, and $\boldsymbol{d}$ is the vector of $(K+3)$ eigenvalue sorted in descending orders, and $\boldsymbol{U}\boldsymbol{U}^T = I$. Further, they decompose $\boldsymbol{d} = (\boldsymbol{d}_+^T, \boldsymbol{d}_0^T)^T$, where $\boldsymbol{d}_+^T$ is vector of $(K+2)$ descending positive eigenvalues, and d_0^T is the zero eigenvalue with one element. Additionally, $\boldsymbol{U} = (\boldsymbol{U}_+, \boldsymbol{U}_0)$, where $\boldsymbol{U}_+$ is a matrix of $(K+3)\times(K+2)$ corresponding to $\boldsymbol{d}_+^T$, and $\boldsymbol{U}_0^T$ is a vector of $(K+3)$ corresponding to $\boldsymbol{d}_0^T$.

Under the above spectral decomposing technique, I partition the B-spline functions as follows:

$$\begin{aligned}\boldsymbol{B}\tilde{\boldsymbol{\beta}} = \boldsymbol{B}I\tilde{\boldsymbol{\beta}} &= \boldsymbol{B}\boldsymbol{U}_\beta\boldsymbol{U}_\beta^T\tilde{\boldsymbol{\beta}} = \boldsymbol{B}\{\boldsymbol{U}_{0\beta}\boldsymbol{U}_{0\beta}^T + \boldsymbol{U}_{+\beta}diag(\boldsymbol{d}_+^{-1/2})diag(\boldsymbol{d}_+^{1/2})\boldsymbol{U}_{+\beta}^T\}\tilde{\boldsymbol{\beta}} \\ &= \boldsymbol{M}_\beta\beta_{lin} + \boldsymbol{N}_\beta\boldsymbol{\beta}_{Non.lin}, \end{aligned} \tag{3.4}$$

and

$$\begin{aligned}\boldsymbol{B}\tilde{\boldsymbol{\gamma}} = \boldsymbol{B}I\tilde{\boldsymbol{\gamma}} &= \boldsymbol{B}\boldsymbol{U}_\gamma\boldsymbol{U}_\gamma^T\tilde{\boldsymbol{\gamma}} = \boldsymbol{B}\{\boldsymbol{U}_{0\gamma}\boldsymbol{U}_{0\gamma}^T + \boldsymbol{U}_{+\gamma}diag(\boldsymbol{d}_+^{-1/2})diag(\boldsymbol{d}_+^{1/2})\boldsymbol{U}_{+\gamma}^T\}\tilde{\boldsymbol{\gamma}} \\ &= \boldsymbol{M}_\gamma\gamma_{lin} + \boldsymbol{N}_\gamma\boldsymbol{\gamma}_{Non.lin}, \end{aligned} \tag{3.5}$$

where $\boldsymbol{M}_{\boldsymbol{\beta}} = \boldsymbol{B}\boldsymbol{U}_{0\beta}$, $\boldsymbol{\beta}_{lin} = \boldsymbol{U}_{0\beta}^T\tilde{\boldsymbol{\beta}}$, $\boldsymbol{N}_\beta = B\boldsymbol{U}_{+\beta}diag(\boldsymbol{d}_+^{-1/2})$, and $\boldsymbol{\beta}_{Non.lin} = diag(\boldsymbol{d}_+^{1/2})\boldsymbol{U}_{+\beta}^T\tilde{\boldsymbol{\beta}}$. Similarly, $\boldsymbol{M}_\gamma = B\boldsymbol{U}_{0\gamma}$, $\boldsymbol{\gamma}_{lin} = \boldsymbol{U}_{0\gamma}^T\tilde{\boldsymbol{\gamma}}$, $\boldsymbol{N}_\gamma = B\boldsymbol{U}_{+\gamma}diag(\boldsymbol{d}_+^{-1/2})$, and $\boldsymbol{\gamma}_{Non.lin} = diag(\boldsymbol{d}_+^{1/2})\boldsymbol{U}_{+\gamma}^T\tilde{\boldsymbol{\gamma}}$.

Matrix $\boldsymbol{M}$ is the $n\times 1$ design matrix for the linear part, and matrix $\boldsymbol{N}$ is the $n\times(K+2)$ design matrix for nonlinear part. Herein, for the logistic regression model, $\boldsymbol{\beta}_{lin}$ and $\boldsymbol{\beta}_{Non.lin}^T$ are the respective coefficient vectors of the linear and nonlinear components with one element and $(K+2)$ elements. Similarly in the proportional

hazard model, $\boldsymbol{\gamma}_{lin}$ and $\boldsymbol{\gamma}^T_{Non.lin}$ are the respective coefficient vectors of the linear and nonlinear components with one element and $(K+2)$ elements. With such a decomposition, the design matrix for logistic model expands to $p(K+3)$ columns, in which p columns correspond to linear elements, and $p(K+2)$ columns correspond to nonlinear elements. In the proportional hazard model, q columns of the design matrix correspond to the linear components, and $q(K+2)$ correspond to the nonlinear components. Thus, for design matrix the β^T_{lin} and γ^T_{lin} are coefficients vectors of p elements correspond to linear parts and the $\beta^T_{Non.lin}$ and $\gamma^T_{Non.lin}$ are coefficients vector of $(K+2)$ elements correspond to nonlinear parts.

By separating the functional forms of independent variables into linear and nonlinear parts, I am able to give different penalties to the two parts. Using Equations (3.4, 3.5), I rewrite the log-likelihood function (3.3) as follows:

$$
\begin{aligned}
l_C(\boldsymbol{\beta}_{lin}, \boldsymbol{\beta}_{Non.lin}, \boldsymbol{\gamma}_{lin}, \boldsymbol{\gamma}_{Non.lin}, \lambda_{uc,0}; y) = & \\
& \sum_{i=1}^{n} \Big\{ y_i (\boldsymbol{M}_{\beta,i}\boldsymbol{\beta}_{lin} + \boldsymbol{N}_{\beta,i}\beta_{Non.lin}) \\
& - log\{1 + \exp(\boldsymbol{M}_{\beta,i}\boldsymbol{\beta}_{lin} + \boldsymbol{N}_{\beta,i}\boldsymbol{\beta}_{Non.lin})\} \Big\} \\
& + \sum_{i=1}^{n} \Big\{ y_i \delta_i (log\lambda_{nc,0}(t_i) + \boldsymbol{M}_{\gamma,i}\boldsymbol{\gamma}_{lin} + \boldsymbol{N}_{\gamma,i}\boldsymbol{\gamma}_{Non.lin}) \\
& - y_i e^{\boldsymbol{M}_{\gamma,i}\boldsymbol{\gamma}_{lin} + \boldsymbol{N}_{\gamma,i}\boldsymbol{\gamma}_{Non.lin}} \Lambda_{nc,0}(t_i) \Big\}.
\end{aligned}
$$

For selection, I use an adaptive LASSO procedure for which the penalty terms have a sparsity condition, small coefficients to zero, on solutions for β_{lin}, $\beta_{Non.lin}$, γ_{lin}, and

$\gamma_{Non.lin}$. For solution, I maximize

$$
\begin{aligned}
&pl_c(\boldsymbol{\beta}_{lin}, \boldsymbol{\beta}_{Non.lin}, \boldsymbol{\gamma}_{lin}, \boldsymbol{\gamma}_{Non.lin}, \lambda_{nc,0}; y) \\
&\quad = \{l_{C,1}(\boldsymbol{\beta}_{lin}, \boldsymbol{\beta}_{Non.lin}; y) - \tau_{1,lin} \sum_{j_1=1}^{p} \frac{|\boldsymbol{\beta}_{lin,j_1}|}{|W_{lin,j_1}|} - \tau_{1,Non.lin} \sum_{j_2=1}^{p} \frac{|\boldsymbol{\beta}_{Non.lin,j_2}|}{|W_{Non.lin,j_2}|}\} + \\
&\{l_{C,2}(\boldsymbol{\gamma}_{lin}, \boldsymbol{\gamma}_{Non.lin}, \lambda_{nc,0}; y) - \tau_{2,lin} \sum_{k_1=1}^{q} \frac{|\boldsymbol{\gamma}_{lin,k_1}|}{|C_{lin,k_1}|} - \tau_{2,Non.lin} \sum_{k_2=1}^{q} \frac{|\boldsymbol{\gamma}_{Non.lin,k_2}|}{|C_{Non.lin,k_2}|}\}. \qquad (3.6)
\end{aligned}
$$

where $\tau_{1,lin}$, $\tau_{1,Non.lin}$, $\tau_{2,lin}$, and $\tau_{2,Non.lin}$ are the tuning parameters. I use the Bayesian Information Criteria (BIC) to select tuning parameters (Schwarz, 1978). I discuss the selection of the tuning parameters later in the section.

Weigh functions (W_{lin}, $W_{Non.lin}$, C_{lin}, $C_{Non.lin}$) are consistent estimators of ($\boldsymbol{\beta}_{lin}$, $\boldsymbol{\beta}_{Non.lin}$, $\boldsymbol{\gamma}_{lin}$, $\boldsymbol{\gamma}_{Non.lin}$) (Zou, 2006). I impose different weights on different components. Less important components receive greater weights, with increased penalties. The unimportant functional components are then set to zero and removed from the model. When weights are set to 1, LASSO estimators ensue. The proposed method adds to the existing selection procedures an ability to discriminate zero, linear, and nonlinear effects (Zhang *et al.*, 2011).

I summarize the simultaneous variable selection and structural discovery procedure as follows.

1. Linear variable selection: When $\boldsymbol{\beta}_{lin} \neq 0$ and $\boldsymbol{\beta}_{Non.lin} = 0$ for the logistic model, and $\boldsymbol{\gamma}_{lin} \neq 0$ and $\boldsymbol{\gamma}_{Non.lin} = 0$ for the proportional hazard model.
2. Nonlinear variable selection: When $\boldsymbol{\beta}_{lin} = 0$ and $\boldsymbol{\beta}_{Non.lin} \neq 0$ for the logistic model, and $\boldsymbol{\gamma}_{lin} = 0$ and $\boldsymbol{\gamma}_{Non.lin} \neq 0$ for the proportional hazard model.
3. Sparsity estimation: When $\boldsymbol{\beta}_{lin} = 0$ and $\boldsymbol{\beta}_{Non.lin} = 0$ for the logistic model, and $\boldsymbol{\gamma}_{lin} = 0$ and $\boldsymbol{\gamma}_{Non.lin} = 0$ for the proportional hazard model.

3.2.4 Computation

When a subject is censored, his/her cure status will be unknown. I therefore apply an EM algorithm for computation. I use quadratic approximation algorithm (Fan and Li, 2001) to compute the maximum likelihood estimate of $(\boldsymbol{\beta}_{lin}, \boldsymbol{\beta}_{Non.lin}, \boldsymbol{\gamma}_{lin}, \boldsymbol{\gamma}_{Non.lin})$ as follows.

The E-step: Let $(\boldsymbol{\beta}_{lin}^{(m)}, \boldsymbol{\beta}_{Non.lin}^{(m)}, \boldsymbol{\gamma}_{lin}^{(m)}, \boldsymbol{\gamma}_{Non.lin}^{(m)}, \lambda_{nc,0}^{(m)})$ be the parameter estimates in the mth iteration. Given the observed data $(t_i, \delta_i, \boldsymbol{x_i}, \boldsymbol{z_i})$, in the $(m+1)$th iteration, I replace y_i with the $y_i^{(m+1)}$

$$y_i^{(m+1)} = \delta_i + (1-\delta_i)\frac{e^{\{(\boldsymbol{M}_\beta \boldsymbol{\beta}_{lin}^{(m)} + \boldsymbol{N}_\beta \boldsymbol{\beta}_{Non.lin}^{(m)}) - \Lambda_{nc,0}^{(m)}(t_i)\exp(\boldsymbol{M}_\gamma \boldsymbol{\gamma}_{lin}^{(m)} + \boldsymbol{N}_\gamma \boldsymbol{\gamma}_{Non.lin}^{(m)})\}}}{1 + e^{\{(\boldsymbol{M}_\beta \boldsymbol{\beta}_{lin}^{(m)} + \boldsymbol{N}_\beta \boldsymbol{\beta}_{Non.lin}^{(m)}) - \Lambda_{nc,0}^{(m)}(t_i) exp(\boldsymbol{M}_\gamma \boldsymbol{\gamma}_{lin}^{(m)} + \boldsymbol{N}_\gamma \boldsymbol{\gamma}_{Non.lin}^{(m)})\}}}.$$

The M-step: With $y^{(m+1)}$ plugged in, I maximize (3.6) with respect to (β_{lin}, $\beta_{Non.lin}, \gamma_{lin}, \gamma_{Non.lin}, \lambda_{nc,0}$). The M-step involves the following sub-steps:

1. I estimate the cumulative baseline hazard function $\Lambda_{nc,0}(t)$ with a Breslow-type estimator (Klein, 1982; Sy and Taylor, 2000). The nonparametric estimate in the $(m+1)$th iteration is

$$\Lambda_{nc,0}^{(m+1)}(t) = \sum_{t_l \le t} \frac{d_l}{\sum_{k^* \in R_l} y_{k^*}^{(m+1)} exp(\boldsymbol{M}_{\gamma,k^*} \boldsymbol{\gamma}_{lin} + \boldsymbol{N}_{\gamma,k^*} \boldsymbol{\gamma}_{Non.lin})},$$

   where d_l is the number of events at the earliest time t_l, and R_l is the number of subjects at risk at t_l.

2. In the logistic regression model, I solve the penalized score equations for $\beta_{lin}^{(m+1)}$ and $\beta_{Non.lin}^{(m+1)}$

$$\begin{aligned} U(\boldsymbol{\beta}_{lin}) &= \sum_{i=1}^{n} \boldsymbol{M}_{\beta,i}^{T}\Big[y_i^{(m+1)T} - \theta(x_i)\Big]^T \\ &\quad -\tau_{1,lin}\{diag(\frac{1/|\boldsymbol{\beta}_{lin,j_1}^{(m)}|}{|W_{lin,j_1}|})\}\boldsymbol{\beta}_{lin} = 0, \\ U(\boldsymbol{\beta}_{Non.lin}) &= \sum_{i=1}^{n} \boldsymbol{N}_{\beta,i}^{T}\Big[y_i^{(m+1)T} - \theta(x_i)\Big] \\ &\quad -\tau_{1,Non.lin}\{diag(\frac{1/|\boldsymbol{\beta}_{Non.lin,j_2}^{(m)}|}{|W_{Non.lin,j_2}|})\}\boldsymbol{\beta}_{Non.lin} = 0, \end{aligned}$$

for $j_1 = 1, 2, \ldots, p$, and $j_2 = 1, 2, \ldots, p(K+2)$.

The respective penalized Hessian matrices for $\boldsymbol{\beta}_{lin}$ and $\boldsymbol{\beta}_{Non.lin}$ in $(m+1)$th iteration is given by $H_{\beta_{lin}} = \frac{\partial U(\boldsymbol{\beta}_{lin})}{\partial \boldsymbol{\beta}_{lin}}$, and $H_{\beta_{Non.lin}} = \frac{\partial U(\boldsymbol{\beta}_{Non.lin})}{\partial \boldsymbol{\beta}_{Non.lin}}$.

3. With $\Lambda_{nc,0}^{(m+1)}(t)$, I respectively solve following penalized score equations with respect to $\boldsymbol{\gamma}_{lin}$ and $\boldsymbol{\gamma}_{Non.lin}$,

$$\begin{aligned} U(\boldsymbol{\gamma}_{lin}) &= \sum_{i=1}^{n} \boldsymbol{M}_{\gamma,i}^{T}\Big[y_i^{(m+1)}(\delta_i - \Lambda_{nc,0}^{(m+1)}(t_i)\mu(z_i)^T)\Big]^T \\ &\quad -\tau_{2,lin}\{diag(\frac{1/|\boldsymbol{\gamma}_{lin,k_1}^{(m)}|}{|C_{lin,k_1}|})\}\boldsymbol{\gamma}_{lin} = 0, \\ U(\boldsymbol{\gamma}_{Non.lin}) &= \sum_{i=1}^{n} \boldsymbol{N}_{\gamma,i}^{T}\Big[y_i^{(m+1)}(\delta_i - \Lambda_{nc,0}^{(m+1)}(t_i)\mu(z_i)^T)\Big]^T \\ &\quad -\tau_{2,Non.lin}\{diag(\frac{1/|\boldsymbol{\gamma}_{Non.lin,k_2}^{(m)}|}{|C_{Non.lin,k_2}|})\}\boldsymbol{\gamma}_{Non.lin} = 0, \end{aligned}$$

for $k_1 = 1, 2, \ldots, q$, and $k_2 = 1, 2, \ldots, q(K+2)$.

The respective penalized Hessian matrix for $\boldsymbol{\gamma}_{lin}$ and $\boldsymbol{\gamma}_{Non.lin}$ in $(m+1)$th iteration is given by $H_{\gamma_{lin}} = \frac{\partial U(\boldsymbol{\gamma}_{lin})}{\partial \boldsymbol{\gamma}_{lin}}$, and $H_{\gamma_{Non.lin}} = \frac{\partial U(\boldsymbol{\gamma}_{Non.lin})}{\partial \boldsymbol{\gamma}_{Non.lin}}$.

3.3 Tuning Parameter Selection

In the penalized function (3.6), I have tuning parameters τ= ($\tau_{1,lin}$, $\tau_{1,Non.lin}$, $\tau_{2,lin}$, and $\tau_{2,Non.lin}$). These tuning parameters serve an essential function in penalized likelihood estimation. As τ increases, more coefficients shrink to zero (Zou, 2006), thus increasing the bias (Zhang and Lu, 2007). Nishii (1984) proposed a generalized information criterion (GIC) to select the τ. A GIC-type regularization parameter selector has the form

$$GIC(\tau) = \frac{1}{n}\{l_C + \kappa df_\tau\}, \tag{3.7}$$

where df_τ is the degrees of freedom of model (3.6).

I select the combination of $\tau_{1,\beta_{lin}}$, $\tau_{1,Non.lin}$, $\tau_{2,lin}$, and $\tau_{2,Non.lin}$ that minimizes equation (3.7) for a given κ. As κ increases, the size of selected model decreases. When $\kappa = \log(n)$, the GIC type selector reduces to the traditional BIC selector (Schwarz, 1978). The BIC selector not only identifies the true model consistently (Zou and Li, 2008), but is also asymptotically efficient (Zhang *et al.*, 2010). Thus, I use the BIC selector to obtain solutions for $\boldsymbol{\beta}_{lin}$, $\boldsymbol{\beta}_{Non.lin}$, $\boldsymbol{\gamma}_{lin}$ and $\boldsymbol{\gamma}_{Non.lin}$.

3.4 Post-Selection Inference

Finally, I briefly comment on the post-selection parameter estimation and inference. Making valid inference in the selected models poses a new set of challenges. In general, the L_1 penalty-introduced biases affect both estimation and inference. A convenient approach to minimize the biases is to refit the selected model without the penalty terms. Such a two-stage approach is consistent with the practice where inferences are based on selected models (Moore and McCabe, 2009). I follow this two-stage model fitting process in this research. I note, however, Berk and colleagues (2013) have placed the post-selection inference of linear models in the framework of multiple comparison, in which they account for the errors associated with all possible sub-

models. While the idea is intuitively appealing, its validity in nonlinear models needs further investigation.

Following the standard two-stage approach, I use a resample method to construct confidence intervals for the selected model parameters. Briefly, in the first stage, I select variables and determine effect structures, with adaptive LASSO penalties. In the second-stage, I refit the selected model and obtain the estimates for the selected effects without penalties. For inference, I use bootstrap to obtain 95%confidence intervals for the selected effects.

3.5 Simulation Study

I conducted a simulation study to evaluate numerical characteristics of the proposed procedure under LASSO and adaptive LASSO penalties.

For data generation, I assumed that approximately 30% of the subjects are cured. I used following logistic model to generate the cured subjects

$$\begin{aligned}\log\{1-\frac{\theta(x_i)}{1-\theta(x_i)}\} &= 2.5x_{i,1}+0x_{i,2}+1.5x_{i,3}+0.2x_{i,4}+h_5(x_{i,5})+h_6(x_{i,6})\\ &\quad -0.5x_{i,7}+h_8(x_{i,8})+0x_{i,9}+0x_{i,10},\end{aligned}$$

where x_1, x_2 followed binomial distributions with event probabilities set to 0.5. Variables x_3, x_9 and x_{10} followed a multivariate standard normal distribution with a pairwise correlation $0.5^{|i-j|}$. Variable x_4 followed a uniform(0,1) distribution. Variable x_7 followed a multinomial distribution that took values: $0, 1, 2, 3$, with equal probabilities of 0.25. The other variables had nonlinear effects, $h_5(x) = \sin(2\pi x); h_6(x) = \cos(2\pi x); h_8(x) = (3x-1)^2$. The domains of these functions were $x \in (0, 1)$.

For the non-cured subjects, I generated failure times from a Weibull distribution with survival function $S(t|a, b) = exp\{-\frac{t}{b}\}^a$. The shape parameter was set to $a = 1.5$, and scale parameter was set to $b = exp\{e^{\mu(x_i)}\}^{-1/a}$, where $\mu(z_i)$ had the following

additive form

$$\log(\mu(z_i)) = 0z_{i,1}+2.5z_{i,2}+0z_{i,3}+0z_{i,4}+h_5(z_{i,5})+h_6(z_{i,6})+1.5z_{i,7}+h_8(x_{i,8})+0z_{i,9}+0z_{i,10}.$$

Without loss of generality, I assumed $\boldsymbol{x} = \boldsymbol{z}$ in the simulation. I generated the censoring times from exponential distribution with mean of d^{-1}, where d was selected to achieve the desired censoring rate. I considered two different rates of censoring: 40% and 60%. For each parameter setting, I generated 100 datasets with two different sample sizes $n = 250$ and $n = 350$.

With generated data, I fit cure rate models and used the proposed method to determine the structure of the model. Nonlinear effects were modeled with B-splines (Eilers and Marx, 1996). As a sensitivity analysis, I evaluated the influence of knots on selection performance, by considering three different numbers of knots, $K = 3$, 6 and 9.

For estimation, I reparameterized the nonlinear functions, $h_5(\cdot)$, $h_6(\cdot)$ and $h_8(\cdot)$. For each K I constructed the penalized log-likelihood function (3.6). I implemented the selection procedure with a Breslow-type estimator for the baseline hazard $\Lambda_{nc,0}(t)$ in the EM algorithm. For the adaptive LASSO method, I maximized the unpenalized log-likelihood, to obtain the adaptive weights. I conducted the simulation studies with different values of tuning parameter $\tau = (\tau_{1,lin},\ \tau_{1,Non.lin},\ \tau_{2,lin},$ and $\tau_{2,Non.lin})$. Optimal values of the tuning parameters were selected by minimizing the BIC selector (3.7). I evaluated the proposed methods based on: (a) the rate of correct selection for the linear variables; and (b) rate of correct identification of the nonlinear effects.

The logistic regression model included seven linear effects (four zero effects and three nonzero effects) and three nonlinear effects. The proportional hazard model similarly included seven linear effects (five zero effects and two nonzero effects), and three nonlinear variables. Table 3.1 presents the selection results of the linear effects.

Herein, I reported the average numbers of correct exclusion (unimportant effects not being selected) and average number of incorrect exclusion (important effects not being selected) for the logistic and the proportional models. The results were based on 100 simulations.

For the logistic regression model, the average number of incorrect exclusions is zero. The average number of correct exclusion approached the true number of zero effect 3 as sample size increased, for both LASSO and adaptive LASSO selection methods. The selection performance showed a similar selection performance when $K = 6$ and 9, regardless of the censoring rates.

For the proportional hazard model, both selection methods correctly selected all non-zero effects, indicating a zero rate of incorrect exclusion. The two selection methods had different rates of correct exclusion. The adaptive LASSO procedure excluded, on average, between 3.18 and 5 of the five true zero effects. The LASSO procedure excluded fewer true zero effects. The current simulation study suggested that both selection methods had decent performance when the sample size was sufficiently large. Selection results were generally robust against varying number of knots.

Simulation results on selection of nonlinear effects are reported in Table 3.2. Here I presented selection accuracy as proportions of nonlinear effects identified. In analysis, if one or more nonlinear coefficients were selected as nonzero, variable was considered as nonlinear. Simulation results showed that both LASSO and adaptive LASSO had a similar level of selection accuracy in the logistic and proportional hazard models across all parameter settings, suggesting that both procedures were capable of identifying the true underline structures of nonlinear variable effects. The adaptive LASSO method tended to outperform the LASSO method with increasing sample size. Numbers of knots did not significantly influence the selection accuracy in identification of nonlinear effects.

I further investigated the inference performance of the proposed two-stage method, where inference was made in the second stage of model fitting, in the absence of penalty. Table 3.3 presents the average width and the coverage probabilities of the bootstrap 95% confidence intervals. In addition to showing the performance for the nonzero effects of the linear variables, I report the integrated square error $ISE = E\{h(.) - \hat{h(.)}\}^2$ of the selected nonlinear variables. The ISE is calculated via a simulation over 100 replications. Table 3.4 presents a summary result of the replication. I also present the estimated functional curves from the logistic model and the PH model. Figure 3.1 and Figure 3.2 plot the estimated curves in one arbitrary realization of simulated dataset. The simulation showed an overall consistent nonlinear curve estimator $\hat{h}(.)$ and generally good coverage for the nonzero effects. Thus, in the absence of theoretical development of post-selection inference, the simulation result suggested that the bootstrap confidence intervals provides an *ad hoc* but reasonable approximate inference in the mixture cure rate models.

3.6 Application

I analyzed data from a study of childhood high blood pressure development using the proposed method. Design and protocol of the study was described elsewhere (Tu *et al.*, 2011). Briefly, it is an observational study aimed at understanding childhood factors associated with blood pressure. Healthy children aged $5 - 17$ were recruited from schools of Indianapolis, Indiana. Participants were followed prospectively, with semi-annual assessments of blood pressure. Blood and overnight urine samples were collected and analyzed. In the current analysis, I considered various clinical characteristics of the study participants and investigated their associations with blood pressure elevation. Because levels of blood pressure tend to increase with age and height, I converted blood pressure into age, sex, and height-adjusted percentile values followed the existing clinical guidelines. Blood pressure measurements in the top 95th

percentile were considered hypertensive (NHBPEP, 2004). I used a mixture cure rate model for analysis because many children were not at risk of blood pressure spikes.

A total of nine variables were considered in the analysis, including sex (FEMALE,1=female and 0=male), race (BLACK, 1=black, 0=otherwise), mother had hypertension during pregnancy (MOMHIGBP, 1=yes and 0=no), and father had hypertension (FHIGBP, 1= yes and 0=no), body mass index (BMI=weight/height^2), heart rate (PULSE), volume of overnight urine per creatinine (Log(UVOLCr)), urinary sodium excretion rate per creatinine (Log(UNaCr)); (5) urinary potassium excretion rate per creatinine (Log(UKCr)). Summary of characteristics of these variables are reported in table 3.5.

Probabilities of blood pressure not being in the hypertensive range at different ages are depicted by the Kaplan-Meier plot in Figure 3.3. The flattened curves in the upper age spectrum suggested that a significant portion of the study participants were free of the risk of blood pressure elevation in the hypertensive range, and thus providing empirical support for the use of cure rate model (Sy and Taylor, 2000). I analyzed the data using the mixture cure rate model in (3.2). I performed variable selection and nonlinear effect discovery using the methods described in this paper. For detection of the nonlinear effects, continuous variable effects were modeled with cubic B-splines with a large number of interior knots. Both LASSO and the adaptive LASSO methods were used.

As previously described, I selected the tuning parameters for the logistic and proportional hazard models by minimizing the BIC selector (3.7). Using the adaptive LASSO penalty, I selected FHIGBP, PULSE, Log(UVOLCr), Log(UKCr), and BMI for the logistic model, as linear effects. LASSO penalty gave the same selection results. For the proportional hazard model, LASSO and the adaptive LASSO selected all variables except MOMHIGBP. The adaptive LASSO identified Log(UVOLCr) as having a nonlinear effect, while the LASSO penalty identified all continuous variables

were nonlinear variables. Based on the simulation results, I opted to use the model selected by the adaptive LASSO method. I refitted the model with all selected variables. I used cubic- B splines to approximate the nonlinear effect of Log(UVOLCr), for the proportional hazard model. I present the final model fitting results in Table 3.6.

The selected logistic regression model showed that a greater BMI significantly increased risk of having elevated blood pressure. The proportional hazard model showed that male sex, paternal hypertension, and higher BMI levels were associated with early incidence of hypertension. After adjusting for the covariate effects, the risk of hypertension in males is almost 4 times of females. A larger overnight urine volume was also associated with increased hypertension risk, although the effect was not linear. See Figure 3.4.

3.7 Discussion

In this research, I attempt to address two essential questions in the building of cure rate models: (1) what independent variables should I include in each of the two modeling components, and (2) whether any continuous independent variables should be included as nonlinear effects. The second issue is especially relevant in biological studies as true linear effects are rare in complex biological systems. Although linear approximation of nonlinear effects is the standard practice, treating nonlinear effects as linear still represents a model misspecification, and incorrect use of the functional forms could lead to questionable estimation and inference. For cure rate models, because of the complicated modeling structure, variable selection and nonlinear effect detection present an added challenge. Herein, I propose a set of model selection tools based on LASSO and adaptive LASSO. Empirical evidence suggested that the proposed procedures resulted respectable selection performance. I also illustrated the use of the methods through a real data application.

As a variable selection and structural discovery tool for an important class of models, the methods discussed in the current paper are all based on the concept of regularization. Although various forms of penalty could be used for model selection, adaptive LASSO adds enhanced selection flexibility and results in improved selection accuracy. Importantly, the same idea can be further extended to accommodate other more complex modeling needs, such as time-varying coefficients, independent variable interactions, joint-influences of multiple independent effects, etc. As shown in the current study, nonlinear effects are readily accommodated. Depending on the specific modeling structure, other splines based could be use if necessary. For example, for bivariate effect surfaces, thin-plate splines offer an improved numerical stability (Li *et al.*, 2015).

Although the current research has focused exclusively on the issue of selection, post-selection inference remains a topic of interest. Here I described a two-stage process, in which inference was carried out in the selected models, without the interference of penalty terms. Although the approach is somewhat *ad hoc*, it nonetheless has been successfully employed by a number of authors, in various modeling settings (Zhang *et al.*, 2011,Yan and Huang, 2012; He *et al.*, 2015). In the context of cure rate models, the approach appeared to have a decent performance. In the absence of rigorous theoretical justifications, I note that the bootstrap confidence intervals remain a practical option, as evidenced by my simulation study.

Tables

Table 3.1: Simulation results. Average number of zero coefficients for linear effects. Correct exclusion represents the average number unimportant variables not being selected, Incorrect exclusion represents the average number of nonzero effects not being selected.

		40% censoring			
		Logistic part		PH part	
n	Method	Ave. No of 0 Coeff.		Ave. No of 0 Coeff.	
		Number of Knots = 3			
		Correct exclusion	Incorrect exclusion	Correct exclusion	Incorrect exclusion
250	Adp. Lasso	2.0	0	4.97	0
	Lasso	2.08	0	3.36	0
350	Adp. Lasso	3	0	5	0
	Lasso	3	0	5	0
		Number of Knots = 6			
250	Adp. Lasso	2.0	0	4.93	0
	Lasso	2.09	0	3.34	0
350	Adp. Lasso	3	0	5	0
	Lasso	3	0	5	0
		Number of Knots = 9			
250	Adp. Lasso	2.0	0	4.95	0
	Lasso	2.20	0	3.45	0
350	Adp. Lasso	3	0	5	0
	Lasso	3	0	5	0
		60% censoring			
		Number of Knots = 3			
250	Adp. Lasso	2.0	0	3.21	0
	Lasso	2.10	0	2.11	0
350	Adp. Lasso	3	0	5	0
	Lasso	3	0	5	0

			Number of Knots = 6		
250	Adp. Lasso	2.0	0	3.39	0
	Lasso	2.12	0	2.12	0
350	Adp. Lasso	3	0	5	0
	Lasso	3	0	5	0
			Number of Knots = 9		
250	Adp. Lasso	2.0	0	3.18	0
	Lasso	2.13	0	2.62	0
350	Adp. Lasso	3	0	5	0
	Lasso	3	0	5	0

Table 3.2: Simulation results. Percent of correct identification of nonlinear effects.

		40% censoring					
n	Method	Logistic part			PH part		
		$h_5(x)$	$h_6(x)$	$h_8(x)$	$h_5(x)$	$h_6(x)$	$h_8(x)$
		Number of Knots = 3					
250	Adp. Lasso	79.77	81.60	73.40	86.82	84.77	89.32
	Lasso	85.20	84.80	81.40	92.40	92.60	95.20
350	Adp. Lasso	93.89	93.26	95.36	97.89	96.84	99.57
	Lasso	94.20	92.20	95.40	99.60	99.00	99.80
		Number of Knots = 6					
250	Adp. Lasso	92.87	92.87	92.62	90.37	89.38	91.87
	Lasso	95.27	95.12	95.37	95.87	94.37	96.37
350	Adp. Lasso	93.89	93.26	95.37	97.89	96.84	99.58
	Lasso	94.20	99.00	99.80	99.60	99.00	99.80
		Number of Knots = 9					
250	Adp. Lasso	98.45	99.18	99.27	96.90	97.54	95.72
	Lasso	97.45	99.00	98.72	95.00	95.90	93.82
350	Adp. Lasso	99.27	99.45	99.45	99.72	100.00	98.73
	Lasso	98.18	98.81	97.72	99.72	99.81	98.82
		60% censoring					
		Number of Knots = 3					
250	Adp. Lasso	85.09	82.54	80.36	87.27	87.27	89.82
	Lasso	81.60	81.60	76.20	89.40	87.80	91.20

350	Adp. Lasso	96.00	94.33	96.33	96.67	95.00	99.66
	Lasso	89.60	90.20	91.20	98.20	98.60	100
				Number of Knots = 6			
250	Adp. Lasso	88.50	88.00	85.75	90.25	90.75	92.00
	Lasso	79.00	79.62	78.37	90.75	91.37	91.12
350	Adp. Lasso	97.37	96.25	98.12	97.62	98.75	99.75
	Lasso	92.37	92.00	92.62	98.00	98.50	99.87
				Number of Knots = 9			
250	Adp. Lasso	85.57	93.87	91.60	95.09	98.02	96.14
	Lasso	84.09	91.90	90.00	91.72	94.90	92.90
350	Adp. Lasso	98.45	99.45	99.54	99.00	100.00	99.63
	Lasso	94.90	98.63	98.45	98.63	99.90	99.72

Table 3.3: Simulation results. CovProb stands for 95% coverage probability, and L stands for the average lengths of 95% bootstrap confidence intervals.

40% censoring &$n = 250$						40% censoring &$n = 350$					
logistic part			PH part			logistic part			PH part		
Variable	L	CovProb	Variable	L	CovProb	Variable	L	CovProb	Variable	L	CovProb
x_1	0.007	0.95	z_2	0.006	0.97	x_1	0.004	0.89	z_2	0.003	0.97
x_3	0.008	0.96	z_7	0.014	0.99	x_3	0.005	0.92	z_7	0.007	0.98
x_4	0.008	0.98				x_4	0.005	0.88			
x_7	0.016	0.96				x_7	0.010	0.90			
60% censoring &$n = 250$						60% censoring &$n = 350$					
x_1	0.011	0.90	z_2	0.007	0.98	x_1	0.007	0.90	z_2	0.004	0.98
x_3	0.013	0.96	z_7	0.012	0.99	x_3	0.008	0.96	z_7	0.007	0.99
x_4	0.014	0.89				x_4	0.009	0.90			
x_7	0.030	0.91				x_7	0.018	0.92			

Table 3.4: Simulation results. Average integrated square errors (and standard errors in parenthesis) for 100 simulations

	n=250 & 40% cen.		n=250 & 60% cen.	
Function	Logistic part	PH part	Logistic part	PH part
h_5	–	0.183	–	0.183
		(0.022)		(0.022)
h_8	0.397	0.394	0.397	0.394
	(0.045)	(0.038)	(0.045)	(0.038)
	n=350 & 40% cen.		n=350 & 60% cen.	
h_5	–	0.180	–	0.0180
		(0.019)		(0.019)
h_8	0.403	0.399	0.403	0.399
	(0.041)	(0.035)	(0.041)	(0.035)

Table 3.5: Baseline characteristics of subjects

Factor		n	Variable	Mean	Variance
FEMALE	0	446	PULSE	83.794	93.057
	1	435			
BLACK	0	556	Log(UVOLCr)	0.406	0.888
	1	325			
MOMHIGBP	0	634	Log(UNaCr)	-2.114	0.286
	1	247			
FHIGBP	0	613	$log(UKCr)$	-3.576	0.318
	1	268			
			BMI	0.316	1.215

Table 3.6: Summary of parameter estimates (95% bootstrap confidence intervals in parentheses). OR stands for odds ratio for the logistic regression model. In the proportional hazard model, HR refers to hazard ratios. $*$ indicates significant variable.

Variable	**OR (CI)**	**HR (CI)**
Intercept	1.90 (0.737, 3.78)	–
FEMALE	–	$0.23^{*}(0.13, 0.36)$
BLACK	–	–
MOMHIGBP	–	–
FHIGBP	1.08 (0.92, 1.29)	$1.60\ ^{*}(1.10, 2.33)$
PULSE	1.00 (0.99, 1.010)	0.99 (0.98, 1.02)
Log(UVOLCr)	1.027 (0.95, 1.09)	
Log(UNaCr)	–	0.96 (0.63, 1.54)
Log(UKCr)	0.98 (0.94, 1.24)	1.11 (0.75, 1.65)
BMI	$1.13^{*}(1.06, 1.24)$	$1.63^{*}(1.34, 2.28)$

Figures

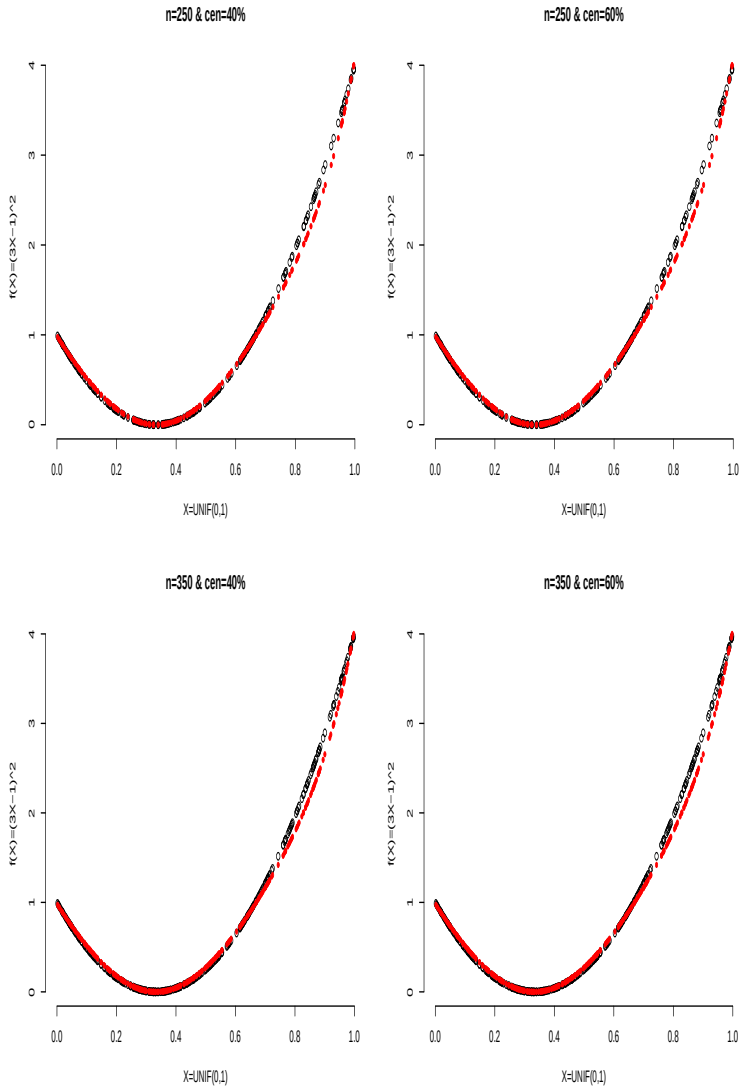


Figure 3.1: Simulation study. Estimated curves from the logistic model; open and black circle represents for true function and closed and red circle represents the estimated function.

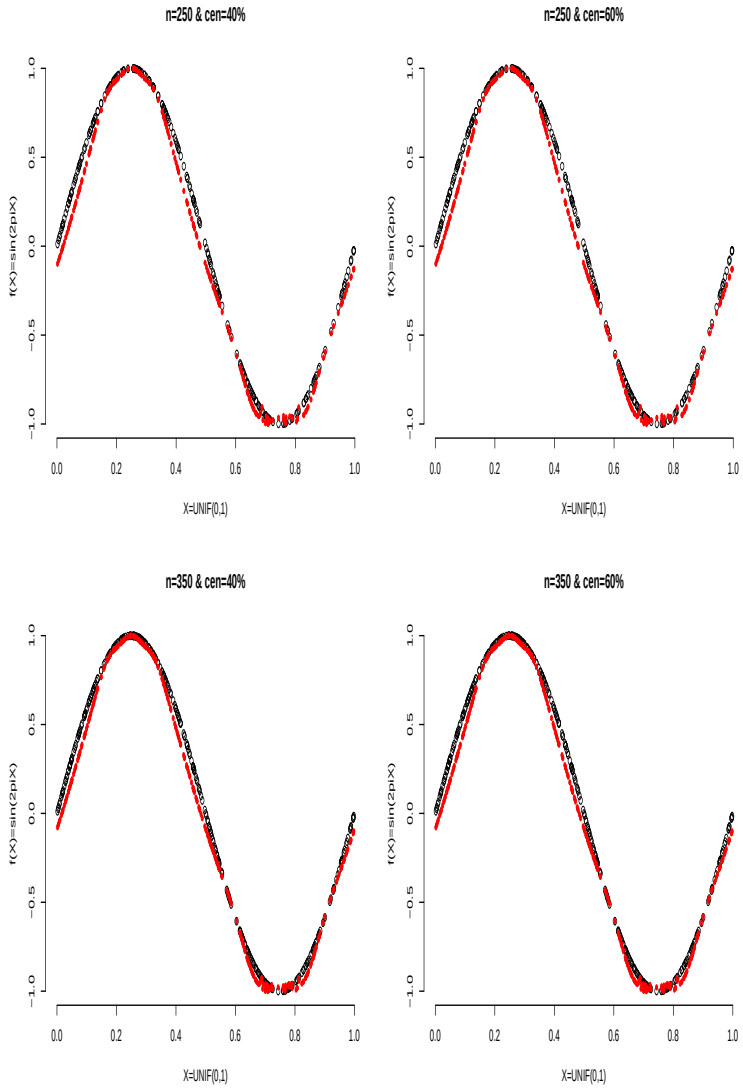


Figure 3.2: Simulation study. Estimated curves from the PH model; open and black circle represents for true function and closed and red circle represents the estimated function.

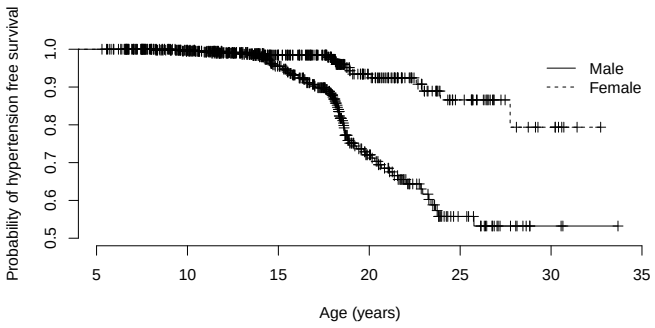


Figure 3.3: Kaplan-Meier curves of male and female subjects

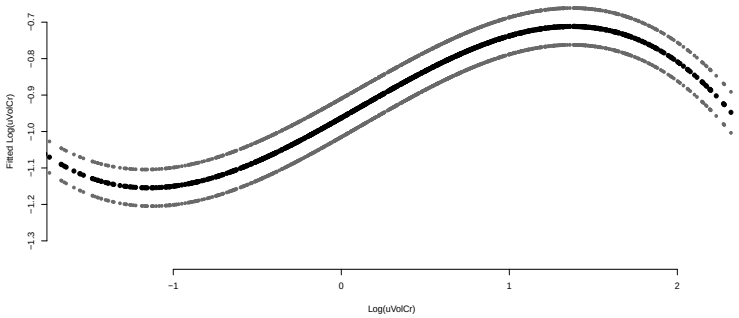


Figure 3.4: Estimated effect of Log(uVolCr) in the proportional hazard model with 95% confidence band.

Chapter 4

VARIABLE SELECTION IN SEMI-PARAMETRIC LINEAR MIXTURE SURVIVAL MODELS FOR CORRELATED FAILURE-TIME DATA

In this chapter, I extend the previously developed variable selection tools to a modeling setting that is similar, but not identical, to the mixture cure rate models. The new modeling context involves repeated failure times, and thus is akin to frailty models; the repeated failures, however, are only observed in the non-cured individuals who *are* at the risk of the event. In other word, the repeated failures do not affect the estimation of the cure probabilities, but they do affect the survival time distribution of the non-cured. This is the modeling setting in which I extend the selection tools. The research is motivated by an epidemiological study of sexually transmitted infections (STI).

4.1 Research Background

The mixture models with two components have grown interest in the analysis of failure-time data. The aim is to assess sub-group of populations who are non-susceptible to the disease of interest while others are highly at risk. Earlier sub-group of population is called long-term survivors, particularly, in cancer clinical trials they are referred to as "cured subjects". Under the conditions presented, interest continues to grow with biomedical applications on cancer studies (Boag, 1949; Berkson and Gage, 1952; Kuk and Chen, 1992; Lambert *et al.*, 2010), studies of hospital readmission data (Yu, 2008; Rondeau, 2010; Rondeau *et al.*, 2011), and studies of smoking cessation (Li *et al.*, 2010). In my motivating biomedical research with adolescent

women where failure event is defined to the occurrence of a certain sexually transmitted infection (STI) such as *Chlamydia trachomatis*, or C *trachomatis*, a certain percentage of women, say $(1 - \Delta)$, may never experience the failure event, which is characterized by the overall Kaplan-Meier (KM) estimate being leveled at non-zero probability. See Figure 4.3(a). This measure quantifies the percentage of long-term survivors of the underlying disease and is a useful to monitor trends in survivorship. The mixture models bridge with the joint models by mixing the survival distribution to describe the long-term survivors. Herein, I refer the mixture models to mixture survival models.

The mixture modeling tool uses a logistic regression formula for Δ and incorporates the Cox regression (Cox, 1972) for the susceptible group. A susceptible subject may have multiple failure times. In the motivating example, a young woman who is at risk of C *trachomatis*, experiences repeated infections. Recurrences after infection involve dependence between times to the failure event. An unobserved random effect, that is called frailty, is used to model the dependency to explain heterogeneous subjects. A common random effect treats multiplicatively on the hazard rates of all susceptible subjects. Frailty models are applied on the correlated data to model recurrent event times such as gap times (kelly and Lim, 2000). The goal of this article is to present model building approach and related model-fitting procedures using Cox frailty model for analyzing the correlated failure times while allowing the logistic regression to depend on a set of baseline independent variables.

Model development is an important task for any regression models. In the current analytical procedure, this task becomes cumbersome as a large number of independent variables are used in composition of the logistic and frailty models. Analysts use a variable selection procedure to obtain the true underlying model that has a spare representation, small coefficients of variables becoming to zero. Fan and Li proposed a penalized likelihood framework to approach the problem of variable selection (Fan

and Li, 2001). Based on the proposed selection method, due to the non-convex form a Smoothly Clipped Absolute Deviation (SCAD) penalty, the resulting estimators are often lack of numerical stability (Zhang and Lu, 2007). Alternatively, Zou (2006) extended Tibshirani's works on the Least Absolute Shrinkage and Selection Operators (LASSO) (Tibshirani, 1996, 1997) to propose an adaptive LASSO method with L_1 penalty, which is computationally stable. In comparison with the LASSO penalty, the adp. LASSO penalty uses data-driven (i.e., adaptive) weights for different coefficients, thus shrinking small coefficients rapidly to zero. In this paper, I propose a regularization method with an adaptive LASSO penalty. Herein, I enhance application of the adaptive LASSO procedure to identify important variables of the mixture survival models for analyzing survival data repeatedly measured over time.

Recent two works on the mixture models with a presence of long-term survivors have been successfully applied the SCAD penalty (Liu *et al.*, 2012) and the adaptive LASSO penalty (Scolas *et al.*, 2016) for variable selection. They assumed the linear relationship between independent and outcome variables. When the linear relationship is violated, the resulting model is at risk of being mis-specified that produces questionable estimates. Additive models alternatively have fewer model assumptions and present much greater modeling flexibility. In survival analysis, these models have been used by many, including Stone (1985) and Hastie and Tibshirani (1990). Literature of model selection with additive effects is limited. Particularly, an additive mixture model for correlated failure times data with the long-term survivors is relatively sparse. The current topic fills the methodology gap in variable selection with nonlinear variables for analyzing complex survival data with an additive modeling setting. My scientific questions of interest are to develop an optimal model: (1) Should a variable included in the model? (2) If so, should it be included as linear or nonlinear effects?

How nonlinear variables should be analyzed, to an extent, depends on the way that these are used in the underlying model. Many literatures are exist for nonlinear regression. Spline based methods, a particular type of nonlinear regression technique, are increasingly used in current analytical practice (Wahba, 1990; Gray, 1992; Wang, 2011). I use B-spline bases for illustrating the nonlinear variables in model formulation because of their optimal stability, which means a change in the coefficients of the bases does not affect in the evaluation of the functional variable (P̃ena, 1997). For Cox regression models, Lin and Halabi (2013) applied an adaptive LASSO penalty to select the functional components. In the current survival data, a cumulative number of partners is often thought to be indicative of C *trachomatis* risk. I can assume an additive effect for the number of partners in the survival model. The additive effect can reduce misspecification of the model. However, correct specification of the functional form of an independent variable can serve a guard for building a valid semiparametric model. Herein, my focus is to construct a useful tool for modeling complex survival data by combining linear and nonlinear effects within one mixture modeling framework.

To detect the structure of the variables, Zhang and colleagues (2011) proposed a data-driven approach by using partially linear models. I extend this approach with variable selection to the mixture models that combine the logistic and Cox frailty models. Under an additive modeling setting, the independent variables have additive effects in the mixture survival models. By partitioning the non-parametric functions into linear component and nonlinear component, I use the L_1 penalty for model selection. This approach is linked to the work by Yan and Huang (2012) who decomposed time-independent and time-dependent coefficients into the Cox model (Cox, 1972). I selected linear effects as time-independent coefficients, and nonlinear effects as time-dependent coefficients. To implement, I use an expectation-maximization (EM) algorithm. Extensive simulations are conducted to evaluate operational char-

acteristics of the proposed methods. Finally, I illustrate the use of the method by analyzing data from a clinical study.

4.2 Model

4.2.1 Formulation

Let T_{ij} is an observed failure time, and δ_{ij} is failure indicator for ith subject, $i = 1, 2, \ldots, n$, at jth recurrence, $j = 1, 2, \ldots, n_i = J$. Failure indicator $\delta_{ij} = 1$ if T_{ij} is observed, and $\delta_{ij} = 0$ if T_{ij} is censored, otherwise.

I define an unobserved binary variable Y_i. $Y_i = 1$ indicates the subject experiences a failure event with the probability of $P(Y_i = 1) = \Delta_i$. Therefore, $(1 - \Delta_i)$ is the fraction of long-term survivors who never experience failure event. The quantity Δ can be specified by a logistic model with baseline independent variables $\mathbf{x_i} \in \mathbb{R}^p$ and is given by

$$log\{\frac{\Delta_i}{1-\Delta_i}\} = \sum_{h=1}^{p} f_h(x_{h,i}),$$

where $f^T(x_i) = (f_1(x_{1,i}) \ldots, f_p(x_{p,i}))$ is a vector of nonparametric function of $\mathbf{x_i}$. I centralize the $\mathbf{x_i}$ to ensure identifiability.

For susceptible subjects with recurrent failure events, suppose $\mathbf{z_{ij}} \in \mathbb{R}^q$ be the vectors of independent variables, which are measured at t_{ij}. t_{ij} is a value of T_{ij}. Conditional on random effect w_i, I assume $t_{i1}, \ldots, t_{ij}$ are independent. The frailty models, which are the *so-called* Cox frailty (CF) models are written as

$$\lambda(t_{ij} = t|z_{ij}, Y_i = 1) = \lambda_0(t) \exp\{\sum_{k=1}^{q} g_k(z_{k,ij}) + w_i\},$$

where $g_k(.)$ is centered and twice differentiable smooth function of kth element in the independent variable vectors $\mathbf{z_{ij}}$. $\lambda_0(t)$ is baseline hazard function. The cumulative

baseline hazard function is

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

w_i is assumed to have an independent normal distribution, $w \sim N(0, a^2)$.

4.2.2 Mixture Survival Models with Random Effects

For ith subject, I define total number of recurrence is $\sum_{j=1}^{J} \delta_{ij}$. When $\sum_{j=1}^{J} \delta_{ij} > 0$, then $Y_i = 1$ with the probability of Δ_i, because the subject experiences at least one failure events. When $\sum_{j=1}^{J} \delta_{ij} = 0$, then the subject is the long-term survivor with the probability of $(1 - \Delta_i)$. The survival probability associated with the long-term survivor is asymptotically (i.e., $t \to \infty$) one. Similar to the susceptible subjects, the long-term survivors have multiple failure times but they correspond to the censored times. Conditional on w_i, the population survival function is

$$\begin{aligned} S_{pop}(t) &= (1 - \Delta_i) + \Delta_i S(t|z_{ij}, Y_i = 1) \\ &= \frac{1}{1 + \sum_{h=1}^{p} f_h(x_{h,i})} + \frac{\sum_{h=1}^{p} f_h(x_{h,i})}{1 + \sum_{h=1}^{p} f_h(x_{h,i})} \times \\ &\quad \exp\Big\{ - \Lambda_0(t|Y_i = 1) \exp(\sum_{k=1}^{q} g_k(z_{k,ij}) + w_i) \Big\} \end{aligned} \tag{4.1}$$

I note here that the above mixture survival models (4.2) are similar to the mixture cure frailty models (Yu, 2008; Rondeau, 2010; Li *et al.*, 2010; Rondeau *et al.*, 2011). The assumption of both models depends on a notion of the underlying event for which a sizable fraction of subjects are censored despite their long follow-up time. Given the background of the censored-survival data, the cured subjects or long-term survivors eventually do not develop the event of interest at end of the study. The presence of those subjects can be empirically observed by the long-tail in the Kaplan-Meier plot (Peng and Taylor, 2014) or can be qualitatively considered by a practical insight (Kuk and Chen, 1992).

Based on a sub-group of population characteristic or a potent treatment condition, in many clinical studies done by Yu (2008), Rondeau *et al.* (2011), and Li *et al.* (2010), the long-term survivors can be treated as the cured subjects. When I use cure models, I strongly assume that the long-term survivors of the event are part of the general population. Therefore, the model assumptions require careful justification (Farewell, 1982). When the KM plot of the time to the event shows a clear level plateau with many censored subjects, the choice of the mixture cure model appeals to analyze the cure fraction in censored-survival data.

Evidence based recommendations about younger age and frequency of concurrent C *trachomatis* infections in adolescent women, my motivational example hypothesizes the postulate of a sub-group of women who are not at risk of infection is a fraction of long-term survivors in the study. Thus, I do not expect the infection-free probabilities are long tail with many censoring events. Since the definition of a long-tern survivor has no episode of recurrent infections, the estimation process of a fraction of long-term survivors does not require random effects. I incorporate random effects for susceptible subjects to multiple infections.

4.3 Method

4.3.1 Likelihood Function

For ith subject, when $\sum_{j=1}^{J} \delta_{ij} = 0$, the Y_i is unobservable because the subject may not be followed up for long enough to occur the event. I construct complete likelihood based on $(t_{ij}, \delta_{ij}, x_i, z_{ij}, y_i)$ that includes the observed data and the unobserved y_i. The marginal complete log-likelihood function is written as:

$$l_c = l_1(.) + l_2(.),$$

where the first term, which corresponds to the logistic model, is

$$l_1(.) = \sum_{i=1}^{n} \{y_i log(\frac{e^{\sum_{h=1}^{p} f_h(x_{h,i})}}{1 + e^{\sum_{h=1}^{p} f_h(x_{h,i})}}) - (1 - y_i) log(1 + e^{\sum_{h=1}^{p} f_h(x_{h,i})})\},$$

and the second term, which corresponds to the CF model, is by using the property of the standard normal variable w_i

$$l_2(.) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_i \{\delta_{ij} (log\lambda_0(t|Y_i = 1) + \sum_{k=1}^{q} g_k(z_{k,ij}) + \frac{a^2}{2}) - \Lambda_0(t|Y_i = 1) e^{\sum_{k=1}^{q} g_k(z_{k,ij}) + \frac{a^2}{2}}\}.$$

Model-fitting of correlated failure times data depends on whether the unknown baseline hazard functions $\lambda_0(t)$ are identical or event-specific among the failure events (Kelly and Lim, 2000). I specify, in this regard, $\lambda_0(t)$ is identical for all recurrent events or failure subjects. The $\lambda_0(t)$ can be described by a nonparametric type estimator (Klein, 1982) or a parametric piece-wise exponential function. I will discuss the estimation procedure for the $\lambda_0(t)$ in the following sub-section.

4.3.2 Regularization Method Using An Adaptive LASSO Penalty

I introduce a semiparametric regression technique, where the functional forms of nonlinear variables are estimated by a cubic B-spline with K interior knots. For any functions, let $f_h(x_i)$ and $g_k(z_{ij})$ are, respectively, the cubic B-splines at the observed points $x_{h,1}, \ldots, x_{h,n}$ for $h = 1, 2, \ldots, p$ and at the observed points $z_{k,1j}, \ldots, z_{k,nj}$ for $k = 1, 2, \ldots, q$ and for each $j = 1, 2, \ldots n_i = J$. With a given K, the approximations are $\mathbf{f(x_i)} = \mathbf{B(x_i)}\tilde{\beta}$, and $\mathbf{g(z_{ij})} = \mathbf{B(z_{ij})}\tilde{\gamma}$. $\mathbf{B}$ is the $n \times (K + 3)$ design matrix for the logistic model and is the $nJ \times (K + 3)$ design matrix for the CF model. $\tilde{\beta}^T$ and $\tilde{\gamma}^T$ are the coefficient vectors of $\mathbf{B(x_i)}$ and $\mathbf{B(z_{ij})}$, respectively.

I use an adaptive type regularization technique via the L_1 penalty to select important variables and to identify potential nonlinear effects. The adaptive LASSO penalty (Zou, 2006) is helpful for variable selection, but it fails to detect the true

underlying functional effects. To remedy, I decompose $f(x)$ and $g(z)$ into the linear and nonlinear components and then use the adaptive LASSO selection procedure to distinguish zero, linear, and nonlinear effects. At the same time, I estimate the regression coefficients.

To illustrate the decomposition of the cubic B-splines, I use $f(x)$ for the logistic model as example. This example is extended to describe the $g(.)$ of multiple measurements for the frailty model. Yan and Huang (2012) applied the B-splines by approximating time-varying coefficients for the Cox models (Cox, 1972). Following the approach I approximate the independent variables. Briefly, I expand $B(\mathbf{x_i})$, in which the first part corresponds to the linear component that has the cubic B-spline basis of order one. Rest of the parts correspond to the nonlinear components that include the cubic B-spline bases of order more than one. Suppose that $\tilde{\beta}^T = (\beta_{lin}^T, \beta_{Non.lin}^T)$. For each $h = 1, 2, \ldots, p$ I partition $B_h(\mathbf{x_i})$ into the linear and nonlinear components, each corresponds to $\tilde{\beta}$. I can write as: $\mathbf{B}\tilde{\beta} = \mathbf{M}_\beta \beta_{lin} + \mathbf{N}_\beta \beta_{Non.lin}$. Matrix $\mathbf{M}_\beta$ is the $n \times 1$ design matrix for the linear component, and matrix $\mathbf{N}_\beta$ is the $n \times (K+2)$ design matrix for the nonlinear component. β_{lin} and $\beta_{Non.lin}$ are respective coefficient vectors of the linear and nonlinear components. Of the $p(K+3)$ columns in the entire design matrix, I have p set of columns that correspond to the linear parts and have $p(K+2)$ set of columns that correspond to the nonlinear parts.

Similarly for the CF model, suppose that $\tilde{\gamma}^T = (\gamma_{lin}^T, \gamma_{Non.lin}^T)$. I can write as: $\mathbf{B}\tilde{\gamma} = \mathbf{M}_\gamma \gamma_{lin} + \mathbf{N}_\gamma \gamma_{Non.lin}$, where matrix $\mathbf{M}_\gamma$ is the $nJ \times 1$ design matrix for the linear component, and matrix $\mathbf{N}_\gamma$ is the $nJ \times (K+2)$ design matrix for the nonlinear component. γ_{lin} and $\gamma_{Non.lin}$ are the respective coefficient vectors of the linear and nonlinear components. In the entire design matrix because of this construction, q columns correspond to the linear parts and rest of $q(K+2)$ correspond to the nonlinear parts.

I am now able to give different penalties to the linear and nonlinear components of the functional forms of independent variables. I rewrite the log-likelihood functions of the logistic and CF models. The log-likelihood of the logistic model is rewritten as

$$l_1(.) = \sum_{i=1}^{n} \Big\{ y_i(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin}) - log\{1 + \exp(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin})\} \Big\}$$

and of the CF model is rewritten as:

$$\begin{aligned} l_2(.) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \Big\{ & y_i\delta_{ij}(log\lambda_0(t|Y_i = 1) + \mathbf{M}_{\gamma,ij}\gamma_{lin} + \mathbf{N}_{\gamma,ij}\gamma_{Non.lin} + \frac{a^2}{2}) \\ & - y_i \exp(\mathbf{M}_{\gamma,ij}\gamma_{lin} + \mathbf{N}_{\gamma,ij}\gamma_{Non.lin} + \frac{a^2}{2})\Lambda_0(t|Y_i = 1) \Big\}. \end{aligned}$$

With these above equations, the complete log-likelihood function of the model (4.2) is rewritten as:

$$\begin{aligned} l_c(.) &= l_1(.) + l_2(.) \\ &= \sum_{i=1}^{n} \Big\{ y_i(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin}) - log\{1 + \exp(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin})\} \Big\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n_i} \Big\{ y_i\delta_{ij}\{log\lambda_0(t|Y_i = 1) + \mathbf{M}_{\gamma,ij}\gamma_{lin} + \mathbf{N}_{\gamma,ij}\gamma_{Non.lin} + \frac{a^2}{2}\} \\ &\quad - y_i \exp(\mathbf{M}_{\gamma,ij}\gamma_{lin} + \mathbf{N}_{\gamma,ij}\gamma_{Non.lin} + \frac{a^2}{2})\Lambda_0(t|Y_i = 1) \Big\}. \end{aligned} \tag{4.2}$$

I propose the following penalized log-likelihood function to estimate the functional forms of independent variables with a sparsity solution, small coefficients to zero, for β_{lin}, $\beta_{Non.lin}$, γ_{lin} and $\gamma_{Non.lin}$:

$$\begin{aligned} pl_c(.) = \{ & l_1(.) - \tau_{1,lin}\sum_{h=1}^{p}\frac{|\beta_{lin,h}|}{|W_{lin,h}|} - \tau_{1,Non.lin}\sum_{h=1}^{p}\frac{|\beta_{Non.lin,h}|}{|W_{Non.lin,h}|}\} \\ & + \{ l_2(.) - \tau_{2,lin}\sum_{k=1}^{q}\frac{|\gamma_{lin,k}|}{|C_{lin,k}|} - \tau_{2,Non.lin}\sum_{k=1}^{q}\frac{|\gamma_{Non.lin,q}|}{|C_{Non.lin,q}|}\}, \end{aligned} \tag{4.3}$$

where $\tau_{1,lin}$, $\tau_{1,Non.lin}$, $\tau_{2,lin}$, and $\tau_{2,Non.lin}$ are the tuning parameters. I use the Bayesian Information Criteria (BIC) (Schwarz, 1978) to select tuning parameters. I continue the discuss of the selection of tuning parameter later in this section. $W_{lin}, W_{Non.lin}, C_{lin}$, and $C_{Non.lin}$ are respective weights for $\beta_{lin}, \beta_{Non.lin}, \gamma_{lin}$, and $\gamma_{Non.lin}$.

Imposing different weights on different coefficients, I enlarge the penalties on uninformative coefficients. If a coefficient is selected to be nonzero, I expect my proposed penalties to make difference between the linear and nonlinear coefficients. The penalized function pl_c is simultaneously used to estimate the regression function and to distinguish zero, linear and nonlinear functions, and thus developing a data-driven model selection tool for the model development. If the weights are set to 1s, I yield LASSO estimators.

I summarize the selection procedure in followings:

1. Obtain the consistent estimates of $\beta_{lin}, \beta_{Non.lin}, \gamma_{lin}$, and $\gamma_{Non.lin}$ maximizing the Equation (4.2), and use the consistent estimates as the adp. weights.
2. Obtain the sparsity solutions if $\beta_{lin} = 0$ and $\gamma_{lin} = 0$ and $\beta_{Non.lin} = 0$ and $\gamma_{Non.lin} = 0$.
3. Obtain the linear effects if $\beta_{lin} \neq$ and $\gamma_{lin} \neq 0$ and $\beta_{Non.lin} = 0$ and $\gamma_{Non.lin} = 0$.
4. Obtain the nonlinear effects if $\beta_{lin} = 0$ and $\gamma_{lin} = 0$ and $\beta_{Non.lin} \neq 0$ and $\gamma_{Non.lin} \neq 0$.

Of note, for the nonlinear components of a functional independent variable, I consider them as a group of components. Thus, any of the components is selected as nonzero the respective independent variable is selected to be nonlinear variable. When the selection procedure selects both linear and nonlinear variables, I fit the optimal model (4.2) by semiparametric regression technique.

4.3.3 Computation

In this section, I derive the EM algorithm to estimate the model parameters. The E-step describes the probability of occurring the failure event. And, the M-step maximizes the log-likelihood function (4.3) with respect to $\beta_{lin}, \beta_{Non.lin}, \gamma_{Lin}, \gamma_{Non.lin}$, and $\lambda_0(.)$ for a given value of a by using the local quadratic approximation algorithm (Fan and Li, 2001).

E-step. Let $(\beta_{lin}^{(m)}, \beta_{Non.lin}^{(m)}, \gamma_{lin}^{(m)}, \gamma_{Non.lin}^{(m)}, \lambda_0^{(m)})$ with a given value of a be the parameter estimates in mth iteration. In $(m+1)$th iteration, I replace y_i in (4.3) with $y_i^{(m+1)}$ for $\delta_i = \sum_{j=1}^{n_i} \delta_{ij}$

$$y_i^{(m+1)} = \delta_i + (1-\delta_i)\frac{e^{(\mathbf{M}_\beta \beta_{lin}^{(m)} + \mathbf{N}_\beta \beta_{Non.lin}^{(m)}) - \Lambda_0^{(m)}(t)\exp(\mathbf{M}_\gamma \gamma_{lin}^{(m)} + \mathbf{N}_\gamma \gamma_{Non.lin}^{(m)} + a^2/2)}}{1 + e^{(\mathbf{M}_\beta \beta_{lin}^{(m)} + \mathbf{N}_\beta \beta_{Non.lin}^{(m)}) - \Lambda_0^{(m)}(t)\exp(\mathbf{M}_\gamma \gamma_{lin}^{(m)} + \mathbf{N}_\gamma \gamma_{Non.lin}^{(m)} + a^2/2)}}.$$

M-step. This step involves the following sub-steps:

1. *The baseline hazard function.* The nonparametric Breslow type estimator for the $\Lambda_0(t)$ has the following form in the $(m+1)$th iteration:

$$\Lambda_0^{(m+1)}(t) = \sum_{t_l \le t} \frac{d_l}{\sum_{k^* \in R_l} y_{k^*}^{(m+1)} \exp(\mathbf{M}_{\gamma,k^*}\gamma_{lin} + \mathbf{N}_{\gamma,k^*}\gamma_{Non.lin} + a^2/2)},$$

   where d_l is the number of failure events at the earliest time t_l, and R_l is the number of subjects at risk at t_l.

2. *The logistic model*

- Score equations for the fixed tuning parameters $\tau_{1,lin}$ and $\tau_{1,Non.lin}$ and for the fixed weights $W_{lin,h}$ and $W_{Non.lin,h}$ for $h = 1, 2, \ldots, p$ are:

$$U(\beta_{lin}) = \sum_{i=1}^{n} \mathbf{M}_{\beta,i}^{T} \Big\{ y_i^{(m+1)} - \frac{\exp(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin})}{1 + \exp(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin})} \Big\}$$
$$-\tau_{1,lin} Diag\{\frac{1/|\beta_{lin,h}^{(m)}|}{|W_{lin,h}|}\}\beta_{lin} = 0,$$

$$U(\beta_{Non.lin}) = \sum_{i=1}^{n} \mathbf{N}_{\beta,i}^{T} \Big\{ y_i^{(m+1)} - \frac{\exp(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin})}{1 + \exp(\mathbf{M}_{\beta,i}\beta_{lin} + \mathbf{N}_{\beta,i}\beta_{Non.lin})} \Big\}$$
$$-\tau_{2,Non.lin} Diag\{\frac{1/|\beta_{Non.lin,h}^{(m)}|}{|W_{Non.lin,h}|}\}\beta_{Non.lin} = 0.$$

- Hessian matrix in the $(m+1)$th iteration is given by $H(\beta_{lin}) = \frac{\partial U(\beta_{lin})}{\partial \beta_{lin}}$ and $H(\beta_{Non.lin}) = \frac{\partial U(\beta_{Non.lin})}{\partial \beta_{Non.lin}}$.

3. *The CF model*

- With given $\Lambda_0^{(m+1)}(t)$ score equations for the fixed tuning parameters $\tau_{2,lin}$ and $\tau_{2,Non.lin}$, for the fixed weights $C_{lin,k}$ and $C_{Non.lin,k}$ for $k = 1, 2, \ldots, q$ and the fixed a are:

$$U(\gamma_{lin}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \mathbf{M}_{\gamma,ij}^{T} \Big\{ y_i^{(m+1)} (\delta_{ij} - \Lambda_0^{(m+1)} e^{\mathbf{M}_{\gamma,ij}\gamma_{lin}^{(m)} + \mathbf{N}_{\gamma,ij}\gamma_{Non.lin}^{(m)} + a^2/2}) \Big\}$$
$$-\tau_{2,lin} Diag\{\frac{|1/\gamma_{lin,k}^{(m)}|}{|C_{lin,k}|}\}\gamma_{lin} = 0,$$

$$U(\gamma_{Non.lin}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \mathbf{N}_{\gamma,ij}^{T} \Big\{ y_i^{(m+1)} (\delta_{ij} - \Lambda_0^{(m+1)} e^{\mathbf{M}_{\gamma,ij}\gamma_{lin}^{(m)} + \mathbf{N}_{\gamma,ij}\gamma_{Non.lin}^{(m)} + a^2/2}) \Big\}$$
$$-\tau_{2,Non.lin} Diag\{\frac{|1/\gamma_{Non.lin,k}^{(m)}|}{|C_{Non.lin,k}|}\}\gamma_{Non.lin} = 0.$$

- Hessian matrix in the $(m+1)$th iteration is given by $H(\gamma_{lin}) = \frac{\partial U(\gamma_{lin})}{\partial \gamma_{lin}}$ and $H(\gamma_{Non.lin}) = \frac{\partial U(\gamma_{Non.lin})}{\partial \gamma_{Non.lin}}$.

The final maximum likelihood (ML) estimate $\hat{\theta} = (\hat{\beta}_{lin},\ \hat{\beta}_{Non.lin},\ \hat{\gamma}_{lin},\ \hat{\gamma}_{Non.lin})$ is achieved by iterating between the E and M steps. In summary, the key steps of the EM algorithms are:

Step 1: Fix the tuning parameter $\tau = (\tau_{1,lin}, \tau_{1,Non.lin}, \tau_{2,lin}, \tau_{2,Non.lin})$ and initialize $\boldsymbol{\theta^{(0)}} = (\boldsymbol{\beta}^{(0)}_{lin}, \boldsymbol{\beta}^{(0)}_{Non.lin}, \boldsymbol{\gamma}^{(0)}_{lin}, \boldsymbol{\gamma}^{(0)}_{Non.lin})$.

Step 2: Execute the E-step and estimate the $\hat{\lambda}_0(t)$.

Step 3: Update the estimates as $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}$ - $H^{-1}(\boldsymbol{\theta}^{(0)})U(\boldsymbol{\theta}^{(0)})$.

Step 4: Repeat step 2 and 3 until $|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}| \to 0$.

4.4 Tuning Parameter Selection

In this section, I discuss a selection criterion for tuning parameter $\tau = (\tau_{1,lin}, \tau_{1,Non.lin}, \tau_{2,lin}$, and $\tau_{2,Non.lin})$, which is used in the penalized log-likelihood (4.3) to simultaneously estimate the regression functions. In the context of penalized variable selection problem, I use the selection criterion that possesses consistent or/and efficient property. Following by Zhang and colleagues (2010), the true model with the consistent property has an asymptotic probability of 1 and with the efficient property yields the lowest mean square error. The BIC (Zou and Li, 2008), the cross-fold validation (CV) (Tibshirani, 1996; Zou, 2006), or the generalized cross-validation (GCV) (Fan and Li, 2001; Zhang and Lu, 2007) type selector has been commonly used to select the tuning parameter in variable selection problem. Nishii (1984) adopted a generalized information criterion (GIC) to select the tuning parameter. The GIC type tuning parameter selector has the form:

$$GIC(\tau) = \frac{1}{n}\{l_c(.) + \kappa df_\tau\}, \tag{4.4}$$

where df_τ is number of parameters used maximizing the function (4.3). When $\kappa = log(n)$, the GIC type selector reduces to a BIC selector (Schwarz, 1978). The

BIC selector identifies the true model consistently (Zou and Li, 2008) and has an asymptotic efficiency (Zhang *et al.*, 2010). Thus, I consider the BIC selector to obtain the solutions for β_{lin}, $\beta_{Non.lin}$, γ_{lin}, and $\gamma_{Non.lin}$ using the EM algorithm. I select τ that minimizes the BIC selector in simulation study and data analysis.

4.5 Post-Selection Inference

In this section, I illustrate a post-selection inference procedure of the selected model. Many selection procedures done by Tibshirani (1996 and 1997), Fan and Li (2001), Zou (2006), Zhang and Lu (2007), and Zou and Li (2008) have been used a full model included all variables to derive asymptotic standard errors for the estimates. In the context of variable selection with structural discovery of the functional independent variables I face nonetheless two fundamental challenges of the semiparametric model. One challenge is model efficiency, with including multivariate nonlinear functions into the logistic and CF models. The model has a large number of dimensions that rapidly increases unacceptable standard errors. To gain the model efficiency, I can refit the selected model and obtain the standard errors of the estimates.

Another challenge is computation and estimation, with executing the EM algorithm with infinite-dimensional function space. The L_1 penalty generates bias in the parameter estimation, thus affecting the inference. To minimize the bias, one can refit the data based on the selected model assuming that the selected model is correct or approximately correct in the finite sample.

These lead us to consider a two-stage approach. This approach has been applied in many complex data analysis. A notable piece of work in this filed is by Zhang and colleagues (2011) who used a two-step procedure to refit data by the selected linear model. In current practice described in the standard textbooks (Moore and McCabe, 2009), the two-stage approach has consistent and common statistical analysis procedures for the selected model to make valid inference. Recently, Berk and

colleagues (2013) who described the post-selection inference in the context of linear regression models by treating the post-selection inference in a multiple hypothesis testing. Each test corresponds with a sub-model. The idea appeals, but its validity in semiparametric regression desires further investigation.

With the nonlinear functions, I use the two-stage approach to improve the model efficiency and to reduce estimation bias. Conditional on the selected variables and the functional structures of the variables, I use the bootstrap method to obtain the standard errors of the estimates. Briefly, suppose that $\hat{\theta}_b$ is the bth bootstrap estimate, $b = 1, 2, \ldots, B$, of a parameter. The estimated bootstrap standard error formula is $\hat{\sigma}_b = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta})^2}$, where $\hat{\theta}$ is an overall estimate the parameter. I summarize the post-selection procedure in the following stages:

1. *First-stage*. Select important variables and identify the functional structures of the variables by using my proposed penalization method.
2. *Second-stage*. This stage has following sub-steps:
   (a) Step 1: *Estimation*. Refit the selected model without the penalty using the derived EM algorithm and obtain the estimates, $\hat{\theta}$, for the logistic and CF models.
   (b) Step 2: *Bootstrap standard error*. Resample data with replacement and execute the EM algorithm based on the selected model. Repeat this process finite B times and obtain the estimated bootstrap standard error, $\hat{\sigma}_b$.

4.6 Simulation Study

I use simulation study to evaluate finite sample performance of the proposed method. Particularly, I compare the rates of selection accuracy between the LASSO and adp. LASSO methods. I perform my simulation study in the **R**-programming language (R core team, 2014).

Data generation. I consider a situation, where the percentage of long-term survivors is 30%. This percentage is described in the following logistic regression:

$$log\{1 - \frac{\Delta(x_i)}{1-\Delta(x_i)}\} = 2x_{i,1} + 0x_{i,2} + 2.5x_{i,3} + 0.2x_{i,4} - 2x_{i,5} + 0x_{i,6} + 0x_{i,7} + 0x_{i,8} + 0x_{i,9} + h_{10}(x_{i,6}) + h_{11}(x_{i,6}) + h_{12}(x_{i,6}).$$

x_1 is binary variable with an equal probability to be 0 or 1, x_2, x_3, x_4 and x_5 are the standard normal variables, x_6, x_7, x_8 and x_9 are the standard uniform variables. I consider these variables as the *linear* variables. The other sets of independent variables have the functional forms, which I call the *nonlinear* variables. Nonlinear variables are: $h_{10}(x) = sin(\pi x)$, $h_{11}(x) = cos(2\pi x)$, and $h_{12}(x) = (1-x)^2$. The domain of each function ranges from 0 to 1. $h_{10}(x)$ and $h_{11}(x)$ are nonlinear function, whereas $h_{12}(x)$ is a partially linear function.

I generate the failure times from the Weibull distribution with a survival function $S(t|a,b) = \exp\{-\frac{t}{b}\}^a$, where $a = 1.5$, and $b = \exp\{exp(\mu)\}^{-1/a}$. μ has the following form:

$$log\{\mu(z_{ij})\} = 0z_{ij,1} + 1.5z_{ij,2} + 0z_{ij,3} + 0z_{ij,4} + 0z_{ij,5} + 0z_{ij,6} - z_{ij,7} + 0z_{ij,8} + 0z_{ij,9} + h_{10}(z_{ij,6}) + h_{11}(z_{ij,6}) + h_{12}(z_{ij,6}).$$

I assume for simplicity that $\mathbf{x_i} = \mathbf{z_i}$. But $\mathbf{z_i}$ has multiple measurements with two, four, and seven repeated occurrences. I generated censor times from $Exp(a)$ distribution, where a is selected to achieve the desired overall censoring rate: 44% and 66%.

For each parameter setting, I generated 100 datasets, with sample sizes of 100, 300 and 600. I consider 5-interior knot and approximate the nonlinear variables from the bs-function with intercept=FALSE in **R**-program (R core team, 2014). I apply

the proposed methods with given values of tuning parameter τ. Optimal values of the tuning parameter are selected by minimizing the BIC selector (4.4).

Simulation results. Table 4.1 presents the selection results of the variables for the mixture model (4.2). The variables with the zero coefficients have no effects in the model. For the logistic model, seven variables have zero effects and rest of five variables have non-zero effects. For the CF model, nine variables have zero effects and the other three variables have non-zero effects. I present the average number of correct exclusion (unimportant variables not being selected) and the average number of correct inclusion (important variables being selected) for the logistic and CF models. The table summarizes the results based on 100 simulations.

Briefly, for both LASSO and adp. LASSO methods, the rate of correct inclusion is 5 for the logistic model. In other words, both penalized methods correctly included all important variables. The adp. LASSO method excludes, on average, 4.80 – 7 of the 7 true unimportant variables. In comparison, the LASSO method excludes, on average, 4.01 – 7 of the 7 true unimportant variables.

For the CF model, both LASSO and adp. LASSO methods have correctly included all 3 important variables. The difference is in the rate of correct exclusion. In this regard, the adp. LASSO has an excellent rate of correct exclusion with a large sample size. It excludes, on average, 3.04 – 9 of the 9 unimportant variables. On the other hand, the LASSO method, on average, can exclude 3.11 – 8.43 of the 9 unimportant variables.

I observe the accuracy (as expressed in percentage) of selecting the functional forms of the nonlinear variables of the logistic and CF models. Table 4.2 reports the performance results under different simulation settings. For the LASSO and adp. LASSO methods, given the nonlinear variables in the logistic and CF model, a large percentage of accuracy of selecting the nonlinear variables is approximately equal.

This is to say that both regularization methods have ability to correctly identify the nonlinear variables.

In comparing the selection performance for the linear variables and the accuracy percentage for identifying the functional structure of the nonlinear variables, the adp. LASSO appears to outperform the LASSO across all simulation settings. Importantly, the performance of the adp. LASSO does not appear to be greatly influenced by censoring percentage or by a large set of repeated occurrences. Overall, the adaptive LASSO has consistent performance across the simulation.

Post-selection inference. I conduct simulation studies under different settings to examine the empirical performance of the statistical inference based on the selected model and the full model. At the first-stage, I identify important variables and obtain their functional structures for related model-fitting by using my proposed adaptive LASSO procedure. I refit the selected model and the full model, then examine the 95% coverage probabilities, the average bootstrap standard error (ASE) and the biases of the nonzero variable coefficients of the logistic and CF models. In calculation of the coverage probability and the ASE, I carry out 100 resample datasets based on the selected model and the full model, respectively. Simulation results are reported in Table 4.3. Briefly, the ASE is approximately equal for the selected and full models. The coverage probabilities are generally good, especially for the selected model, across all parameter settings. The bias is larger for the full model compared with that of the selected model. I also present the estimated functional curves from the logistic model and the frailty model. Figure 4.1 and Figure 4.2 show the estimated curves in an arbitrary realization of simulated dataset for sample size with 600. The simulation showed an overall consistent nonlinear curve estimator $\hat{h}(.)$ and generally good coverage for the nonzero effects. Overall, the simulation results show an empirical support to use bootstrap approach prosed at the second-stage procedure for the post-selected model.

4.7 Application

To illustrate the proposed methods, I consider a real clinical study of sexually transmitted infection (STI). The original study design was outlined elsewhere (Tu *et al.*, 2009; Batteiger *et al.*, 2010). Briefly, it is an epidemiological study of recurrent STI in adolescent women. Subjects between 14 and 17 years of age are recruited in the study and followed up to 54 months. Subjects visited at the local adolescent medicine clinics or health facilities and gave their cervical and vaginal specimens to diagnose for the symptom of *Chlamydia trachomatis*. Subjects were teated at the visit date if the clinicians or nurses found the study infection. Subjects repeatedly visited at the clinics for the diagnosis of the infection every three months. At enrollment, subjects were interviewed with a set of questions related to their lifetime and recent (past two months) sexual behaviors, the age of first sex, the number of sex parters and so forth. I updated theirs answers by each follow up visit.

I aimed to understand the risk factors associated with early onset of C *trachomatis*. For this purpose, variable selection methods that I develop presented a logical tool for risk factor screening. C *trachomatis* was the event outcome. The study data included the gap times (as expressed in year) of the successive occurrences of C *trachomatis* after the enrollment date. The first reinfection time was considered as the time between the enrollment date and the first re-infection time after diagnosed related to the infection. Subject was censored at each gap time if the infection was not found.

The study recruited a total of 387 subjects. A total of 8 risk factors were considered in the current analysis. The only demographic variable was race (Black, 1=black and 0=non-black). The sexual behavior variables in past three moths included: (1) previous infection (Prect, 1=yes and 0=no); (2) infection status at enrollment (Enrollct, 1=yes and 0=no); (3) sexual activity in the past three months (SexAct3Mons, 1=yes and 0=no). Continuous variables included: (1) the age of first sex (Agefrstsex); (2) cumulative number of life partner (nlifePtr); (3) cumulative number of

unprotected sex events in the past three months (nSex3Mons); and (4) cumulative number of partners in the past three months (nPtr3Mons).

Kaplan-Meier (KM) estimates of the infection free overall probabilities and the infection free probabilities by previous infection status are presented, respectively, in Figure 4.3 (a) and 4.3 (b). The KM plots with censoring events, suggesting that a portion of adolescent women were not subject to any risk of C *trachomatis*. To accommodate the portion of the risk-free subjects, I analyzed the data by using the proposed mixture survival models (4.2). I used the standard normal variable as the common random effect parameter that took into account the dependence between the successive events (i.e. gap times) within same subject and reflected the true clinical course of infection in the heterogeneous population. Both LASSO and adaptive LASSO methods were used to select important variables and to identify the underlying functional structures of the selected variables for model development.

I approximated the continuous variables by the cubic B-splines with eight interior knots. For a given set of tuning parameter, I optimized the Equation (4.4) and selected the optimal model that minimized the BIC value. Under both LASSO and adp. LASSO methods, Prect, Black, SexActv3Mons, and nlifePtr were retained for the logistic regression. Also, both methods identified the functional structure of the continuous variable, nlifePtr, was linear. For the Cox frailty model, the adp. LASSO selected 4 variables: Prect, Black, SexActv3Mons, Enrollct, and nPtr3Mons. Whereas, the LASSO selected all variables. The continuous variable, nPtr3Mons, was identified by the adp. LASSO method as a nonlinear variable. Of note, I used a semiparametric regression technique to fit the optimal mixture survival models identified by the adaptive LASSO.

When comparing the results of the selected model, I observed that the estimated coefficients of Prect for the logistic regression and Enrollct for the frailty model were statistically significant. Previous infection status was a risk factor for the infection

with OR=1.45 , CI=(1.04, 2.03). In other words, given previous exposure to infection of C *trachomatis*, the coefficient 1.45 implies the chance of experiencing the infection is approximately $0.60 = \frac{1.45}{2.45}$. The higher risk of the infection was associated with the infection status if detected at enrollment period. HR = 3.52; CI = (2.53, 5.81) means that, in the risk population, for a subject if infection detected at enrollment the risk for C *trachomatis* is approximately 3.52 times no-infection detected at enrollment. I also estimated the functional effects of the nPtr3Mons in the risk of C *trachomatis*. See Figure 4.4, which shows that the risk of infection increases from number 0 to number 4 partners in the past three months and then decreases.

4.8 Discussion

In this article, my fundamental aspects of statistical modeling are important variable selection and model specification with identifying the functional structures of the selected variables. I develop a mixture survival model for analyzing data measured repeatedly over time. The mixture model incorporates a potential fraction of risk-free subjects for survival analysis. I demonstrate my selection procedure and model development process, as well as the related selection performances through extensive simulations. To the best of my knowledge, this is is the first attempt that scientifically investigates such comparisons under a particular type survival data.

My investigation particularly focuses on the steps that produce a flexible model and that reduce misspecification. If the underlying linear assumption of a given model is deviated from the assumption, misspecification gives the biased estimates and results in poor prediction. A correct specification of a variable is often practically not feasible. In the linear regression analysis, I can easily verify the nonlinear relationship between the independent variables and outcome by any graphical tool. However, the lack of any visualization tool challenges identifying the nonlinear functional effects in a complex survival model.

In my selection process, I treat the nonlinear components of a functional independent variable as one group of component. Thus, I apply the same tuning parameter to select the nonlinear components. It opens an interesting question to investigate whether different tuning parameters for different nonlinear components can change the performance. To note that, this effort could challenge obtaining a valid set of tuning parameters because one parameter is used for variable selection and another parameter is used for controlling the smoothness of the function.

An important part of my proposed method suffers from theoretical justification of the post-selection inference. This current work is beyond from the scope of rigorous proof due to my focus on model development with an additive effect. My simulation study supports the notion of a use of resampling technique in a two-stage process and then presents a sensible compromise between bootstrap standard errors based on the full model and the selected model. I have shown in previous works that such two-step procedure works well in joint modeling for longitudinal and survival outcomes (He *et al.*, 2015) and in cure rate modeling settings (Masud *et al.*, 2016).

My model development process and estimation procedure take into account recurrent failure events and involve identifying a group disease free subjects for a given condition of number of total failure $\sum_{j=1}^{n_i} \delta_{ij} = 0$. I could extend the selection procedure by allowing random effects in the logistic model. However, based on the motivating STI example, I aim to quantify the susceptible rate of C *trachomatis* and to describe the risk factors associated with the infection among the heterogeneous population.

My finding on the cumulative number of partners in last three months fewer than five that increases the infection risk, is consist with a previous work done by Yu and colleagues (2012). I elaborately demonstrate estimation procedure via EM algorithm and successfully develop an analytical tool that can all together generalize my pro-

posed method to build a prognostic model used for patient care or any statistical model for analyzing complex data.

Tables

Table 4.1: Simulation study for variable selection. Correct exclusion represents the average number unimportant variables not being selected, Correct inclusion represents the average number of non-zero effects being selected.

		Number of Knots = 5			
		Logistic model		Survival model	
n	Method	Correct exclusion (7)	Correct inclusion (5)	Correct exclusion (9)	Correct inclusion (3)
		Two visits with overall censoring ≈ 44 %			
100	Adp. Lasso	5.40	5	3.82	3
	Lasso	4.96	5	3.41	3
300	Adp. Lasso	6.67	5	5.26	3
	Lasso	6.38	5	4.46	3
600	Adp. Lasso	7.00	5	6.78	3
	Lasso	7.00	5	4.75	3
		Four visits with overall censoring ≈ 44 %			
100	Adp. Lasso	5.40	5	5.15	3
	Lasso	5.46	5	3.10	3
300	Adp. Lasso	6.98	5	5.50	3
	Lasso	7.00	5	3.30	3
600	Adp. Lasso	7.00	5	8.43	3
	Lasso	7.00	5	3.83	3
		Seven visits with overall censoring ≈ 44 %			
100	Adp. Lasso	6.03	5	5.81	3
	Lasso	5.49	5	3.74	3
300	Adp. Lasso	7.00	5	6.14	3
	Lasso	7.00	5	3.82	3

600	Adp. Lasso	7.00	5	9.00	3
	Lasso	7.00	5	7.60	3
			Two visits with overall censoring ≈ 66 %		
100	Adp. Lasso	4.80	5	3.04	3
	Lasso	4.32	5	3.11	3
300	Adp. Lasso	5.08	5	4.10	3
	Lasso	5.15	5	4.84	3
600	Adp. Lasso	7.00	5	4.64	3
	Lasso	7.00	5	4.90	3
			Four visits with overall censoring ≈ 66 %		
100	Adp. Lasso	4.80	5	4.61	3
	Lasso	4.01	5	3.89	3
300	Adp. Lasso	6.92	5	5.44	3
	Lasso	6.95	5	3.54	3
600	Adp. Lasso	7.00	5	8.20	3
	Lasso	7.00	5	3.30	3
			Seven visits with overall censoring ≈ 66 %		
100	Adp. Lasso	4.81	5	5.92	3
	Lasso	3.77	5	3.75	3
300	Adp. Lasso	5.02	5	6.57	3
	Lasso	5.10	5	3.80	3
600	Adp. Lasso	7.00	5	8.75	3
	Lasso	7.00	5	5.30	3

Table 4.2: Simulation study. Accuracy (as expressed percentage) of selecting the nonlinear variables for the logistic and CF models across simulation studies.

Number of Knots = 5			
n	Method	Logistic model	CF model
Two visits with overall censoring ≈ 44%			
		$h_{10}(x)$	$h_{12}(z)$
100	Adp. Lasso	95.46	94.62
	Lasso	87.05	93.61
300	Adp. Lasso	99.38	92.20
	Lasso	91.10	90.02
600	Adp. Lasso	100.00	100.00
	Lasso	100.00	100.00
Four visits with overall censoring ≈ 44%			
		$h_{10}(x)$	$h_{12}(z)$
100	Adp. Lasso	96.30	100.00
	Lasso	91.42	100.00
300	Adp. Lasso	99.71	100.00
	Lasso	92.30	100.00
600	Adp. Lasso	100.00	100.00
	Lasso	100.00	100.00
Seven visits with overall censoring ≈ 44%			
100	Adp. Lasso	96.97	100.00
	Lasso	91.60	100.00
300	Adp. Lasso	100.00	100.00
	Lasso	94.62	100.00
600	Adp. Lasso	100.00	100.00
	Lasso	100.00	100.00

Two visits with overall censoring ≈ 66%			
		$h_{10}(x)$	$h_{12}(z)$
100	Adp. Lasso	93.41	86.06
	Lasso	86.99	83.52
300	Adp. Lasso	98.85	87.28
	Lasso	99.71	95.00
600	Adp. Lasso	100.00	95.43
	Lasso	99.00	99.28
Four visits with overall censoring ≈ 66%			
		$h_{10}(x)$	$h_{12}(z)$
100	Adp. Lasso	96.30	99.50
	Lasso	94.12	99.83
300	Adp. Lasso	97.43	99.85
	Lasso	97.48	99.66
600	Adp. Lasso	100.00	100.00
	Lasso	100.00	100.00
Seven visits with overall censoring ≈ 66%			
100	Adp. Lasso	97.10	100.00
	Lasso	94.87	100.00
300	Adp. Lasso	99.86	100.00
	Lasso	98.43	100.00
600	Adp. Lasso	100.00	100.00
	Lasso	99.71	100.00

Table 4.3: Post selection results. CovP1 and Asd1 stand for coverage probability and average standard deviation based on bootstrap method from the selected model. CovP2 and Asd2 stand for coverage probability and average bootstrap standard deviation based on bootstrap method from the full model. Bias1 represents for absolute value of bias for the selected model, Bias2 represents for absolute value of bias for the full model.

Model	Estimate	CovP1	Asd1	$\lvert Bias1\rvert$	CovP2	Asd2	$\lvert Bias2\rvert$	CovP1	AsdP1	$\lvert Bias1\rvert$	CovP2	Asd2	$\lvert Bias2\rvert$
		Two visits & overall censoring $\approx$ 44%						Two visits & overall censoring $\approx$ 66%					
	$\hat{\beta}_1$	0.98	0.093	0.252	0.93	0.100	0.375	0.96	0.068	0.179	0.92	0.040	0.031
Logistic &	$\hat{\beta}_3$	0.97	0.075	0.395	0.95	0.100	0.542	0.97	0.060	0.265	0.92	0.100	0.391
$n = 100$	$\hat{\beta}_4$	0.95	0.043	0.021	0.94	0.030	0.040	0.97	0.032	0.016	0.95	0.020	0.032
	$\hat{\beta}_5$	0.98	0.064	0.308	0.92	0.040	0.427	0.97	0.049	0.210	0.91	0.040	0.310
PH &	$\hat{\gamma}_2$	0.95	0.020	0.001	0.94	0.040	0.001	0.98	0.038	0.016	0.94	0.010	0.010
$n = 100$	$\hat{\gamma}_7$	0.99	0.010	0.034	1.00	0.030	0.014	1.00	0.026	0.052	1.00	0.010	0.021
		Four visits & overall censoring $\approx$ 44%						Four visits & overall censoring $\approx$ 66%					
	$\hat{\beta}_1$	0.99	0.069	0.486	0.97	0.050	0.518	0.93	0.052	0.394	0.98	0.041	0.402
Logistic &	$\hat{\beta}_3$	0.97	0.057	0.657	1.00	0.032	0.665	0.96	0.045	0.536	0.99	0.025	0.545
$n = 100$	$\hat{\beta}_4$	0.95	0.033	0.042	0.96	0.020	0.047	0.95	0.025	0.040	0.97	0.014	0.045
	$\hat{\beta}_5$	0.96	0.049	0.535	1.00	0.026	0.555	0.93	0.039	0.438	0.98	0.022	0.455

PH &	$\hat{\gamma}_2$	0.92	0.019	0.002	0.97	0.010	0.001	0.98	0.037	0.015	0.99	0.013	0.003
$n = 100$	$\hat{\gamma}_7$	0.86	0.008	0.032	0.99	0.004	0.010	0.95	0.023	0.050	0.99	0.010	0.011
		Seven visits & overall censoring $\approx$ 44%						Seven visits & overall censoring $\approx$ 66%					
	$\hat{\beta}_1$	0.94	0.053	0.037	0.98	0.030	0.513	0.94	0.039	0.021	0.99	0.020	0.453
Logistic &	$\hat{\beta}_3$	0.94	0.044	0.054	1.00	0.020	0.729	0.94	0.034	0.031	0.99	0.020	0.657
$n = 100$	$\hat{\beta}_4$	0.93	0.023	0.002	0.93	0.010	0.045	0.95	0.018	0.003	0.94	0.010	0.045
	$\hat{\beta}_5$	0.96	0.039	0.045	0.99	0.020	0.557	0.95	0.029	0.024	0.99	0.020	0.510
PH &	$\hat{\gamma}_2$	0.95	0.018	0.000	0.98	0.010	0.000	0.95	0.038	0.038	1.00	0.020	0.002
$n = 100$	$\hat{\gamma}_7$	0.98	0.010	0.130	1.00	0.004	0.004	1.00	0.024	0.190	1.00	0.010	0.006
		Two visits & overall censoring $\approx$ 44%						Two visits & overall censoring $\approx$ 66%					
	$\hat{\beta}_1$	0.98	0.049	0.301	0.96	0.050	0.317	0.98	0.036	0.210	0.96	0.030	0.221
Logistic &	$\hat{\beta}_3$	0.97	0.039	0.423	0.94	0.050	0.436	0.99	0.031	0.287	0.96	0.040	0.297
$n = 300$	$\hat{\beta}_4$	0.97	0.023	0.025	0.89	0.030	0.026	0.97	0.016	0.021	0.92	0.020	0.021
	$\hat{\beta}_5$	0.95	0.034	0.340	0.92	0.040	0.348	0.96	0.026	0.230	0.94	0.030	0.237
PH &	$\hat{\gamma}_2$	0.91	0.004	0.000	0.94	0.002	0.000	0.97	0.010	0.006	0.93	0.030	0.001

$n = 300$	$\hat{\gamma}_7$	0.97	0.002	0.011	1.00	0.001	0.002	1.00	0.010	0.018	1.00	0.003	0.004
		Four visits & overall censoring $\approx$ 44%						Four visits & overall censoring $\approx$ 66%					
	$\hat{\beta}_1$	0.96	0.036	0.480	0.99	0.024	0.503	0.95	0.026	0.402	0.98	0.021	0.423
Logistic &	$\hat{\beta}_3$	0.98	0.029	0.677	0.99	0.029	0.681	0.98	0.023	0.561	0.99	0.022	0.566
$n = 300$	$\hat{\beta}_4$	0.95	0.016	0.043	0.94	0.018	0.044	0.96	0.012	0.036	0.94	0.015	0.039
	$\hat{\beta}_5$	0.99	0.025	0.541	0.99	0.024	0.545	0.99	0.019	0.446	0.99	0.019	0.452
PH &	$\hat{\gamma}_2$	0.93	0.004	0.000	0.97	0.002	0.000	0.95	0.007	0.005	0.99	0.004	0.001
$n = 300$	$\hat{\gamma}_7$	0.92	0.002	0.011	1.00	0.001	0.002	0.93	0.005	0.017	1.00	0.002	0.003
		Seven visits & overall censoring $\approx$ 44%						Seven visits & overall censoring $\approx$ 66%					
	$\hat{\beta}_1$	0.97	0.028	0.057	0.99	0.020	0.519	0.92	0.020	0.032	0.99	0.020	0.463
Logistic &	$\hat{\beta}_3$	0.98	0.023	0.091	0.99	0.020	0.694	0.96	0.020	0.050	0.99	0.020	0.620
$n = 300$	$\hat{\beta}_4$	0.99	0.013	0.010	0.95	0.080	0.063	0.95	0.010	0.004	0.95	0.060	0.053
	$\hat{\beta}_5$	0.98	0.019	0.069	0.99	0.020	0.550	0.95	0.015	0.040	0.99	0.010	0.500
PH &	$\hat{\gamma}_2$	0.91	0.003	0.004	0.99	0.003	0.000	0.97	0.007	0.022	1.00	0.005	0.001
$n = 300$	$\hat{\gamma}_7$	0.97	0.002	0.045	1.00	0.001	0.001	0.93	0.008	0.071	1.00	0.003	0.002

		Two visits & overall censoring ≈ 44%						Two visits & overall censoring ≈ 66%					
	$\hat{\beta}_1$	0.95	0.033	0.291	0.93	0.050	0.310	0.95	0.026	0.205	0.94	0.040	0.216
Logistic &	$\hat{\beta}_3$	0.92	0.026	0.407	0.99	0.040	0.413	0.99	0.022	0.281	0.95	0.030	0.286
$n = 600$	$\hat{\beta}_4$	0.96	0.015	0.033	0.95	0.030	0.034	0.95	0.011	0.023	0.96	0.022	0.024
	$\hat{\beta}_5$	0.95	0.023	0.328	0.97	0.033	0.332	0.93	0.018	0.226	0.94	0.033	0.230
PH &	$\hat{\gamma}_2$	0.94	0.001	0.000	0.93	0.001	0.000	0.98	0.005	0.002	0.98	0.002	0.001
$n = 600$	$\hat{\gamma}_7$	0.95	0.001	0.006	1.00	0.001	0.001	1.00	0.003	0.009	1.00	0.001	0.002
		Four visits & overall censoring ≈ 44%						Four visits & overall censoring ≈ 66%					
	$\hat{\beta}_1$	0.98	0.025	0.485	0.99	0.031	0.507	0.96	0.018	0.406	0.99	0.022	0.424
Logistic &	$\hat{\beta}_3$	0.96	0.020	0.662	1.00	0.025	0.666	0.94	0.015	0.552	0.99	0.017	0.553
$n = 600$	$\hat{\beta}_4$	0.97	0.011	0.056	0.98	0.013	0.060	0.97	0.010	0.048	0.99	0.012	0.048
	$\hat{\beta}_5$	1.00	0.020	0.533	0.97	0.020	0.535	0.93	0.014	0.439	0.99	0.015	0.441
PH &	$\hat{\gamma}_2$	0.96	0.001	0.000	0.98	0.001	0.000	0.96	0.002	0.002	0.98	0.002	0.001
$n = 600$	$\hat{\gamma}_7$	0.96	0.001	0.005	1.00	0.001	0.001	0.95	0.002	0.008	1.00	0.001	0.002
		Seven visits & overall censoring ≈ 44%						Seven visits & overall censoring ≈ 66%					

	$\hat{\beta}_1$	0.98	0.020	0.063	0.99	0.020	0.530	1.00	0.014	0.032	0.99	0.015	0.465
Logistic &	$\hat{\beta}_3$	0.96	0.020	0.085	0.99	0.020	0.689	0.98	0.013	0.044	0.99	0.017	0.617
$n = 600$	$\hat{\beta}_4$	0.96	0.010	0.005	0.97	0.090	0.047	0.99	0.010	0.002	0.96	0.010	0.047
	$\hat{\beta}_5$	0.94	0.014	0.067	0.99	0.020	0.552	0.99	0.011	0.035	0.99	0.015	0.491
PH &	$\hat{\gamma}_2$	0.97	0.001	0.002	0.93	0.001	0.000	0.82	0.004	0.010	0.97	0.002	0.000
$n = 600$	$\hat{\gamma}_7$	0.93	0.001	0.023	1.00	0.001	0.001	0.93	0.002	0.032	1.00	0.001	0.001

Table 4.4: Summary of parameter estimates with 95% bootstrap confidence intervals (CI) and two sided p-values for STI study. In the logistic model, OR stands for odds ratio. In the frailty model, HR refers to hazard ratio.

Variable	OR (CI)	p-value	HR (CI)	p-value
Intercept	1.042 (0.709, 1.531)	0.832	–	
Prect	1.451 (1.035, 2.032)	0.030	1.178 (0.880, 1.585)	0.283
Black	1.201 (0.814, 1.772)	0.356	0.961 (0.590, 1.562)	0.876
Enrollct	–		3.520 (2.526, 5.805)	<0.0001
SexActv3Mons	1.220 (0.964, 1.544)	0.096	0.454 (0.035, 5.805)	0.542
Agefrstsex	–		–	
nlifePtr	1.010 (0.961, 1.056)	0.542	–	
nsex3Mons	–		–	
nPtr3Mons	–			

Figures

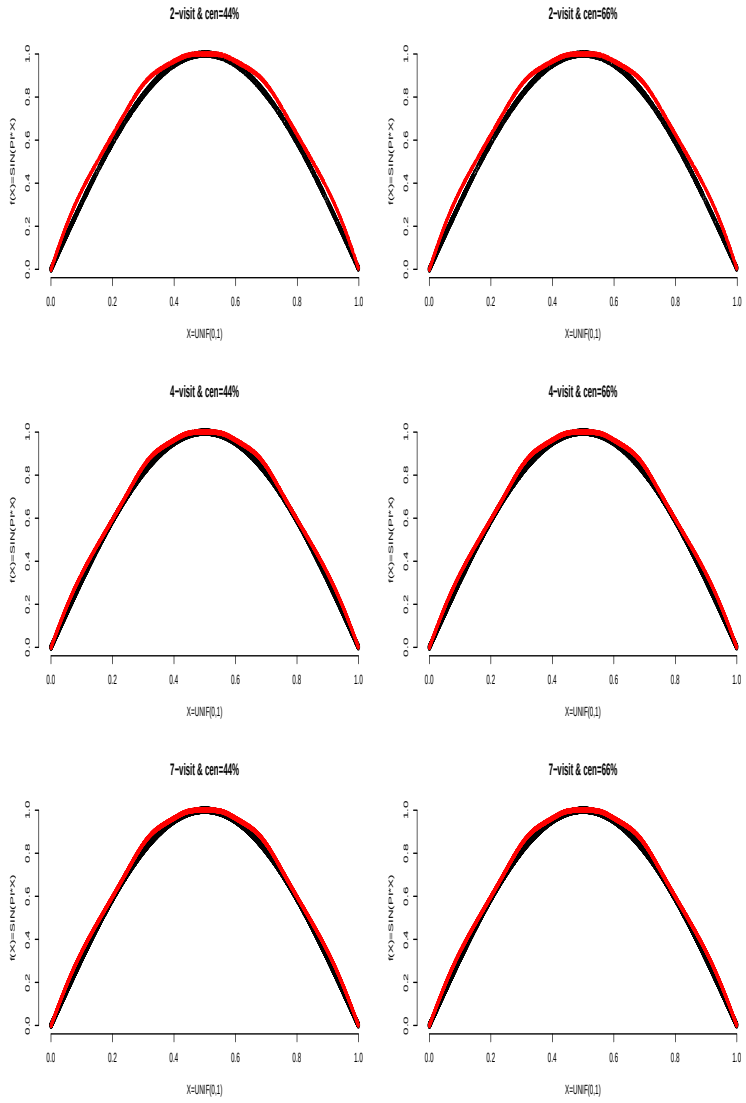


Figure 4.1: Simulation study. Estimated curves from the logistic model; open and black circle represents for true function and closed and red circle represents the estimated function.

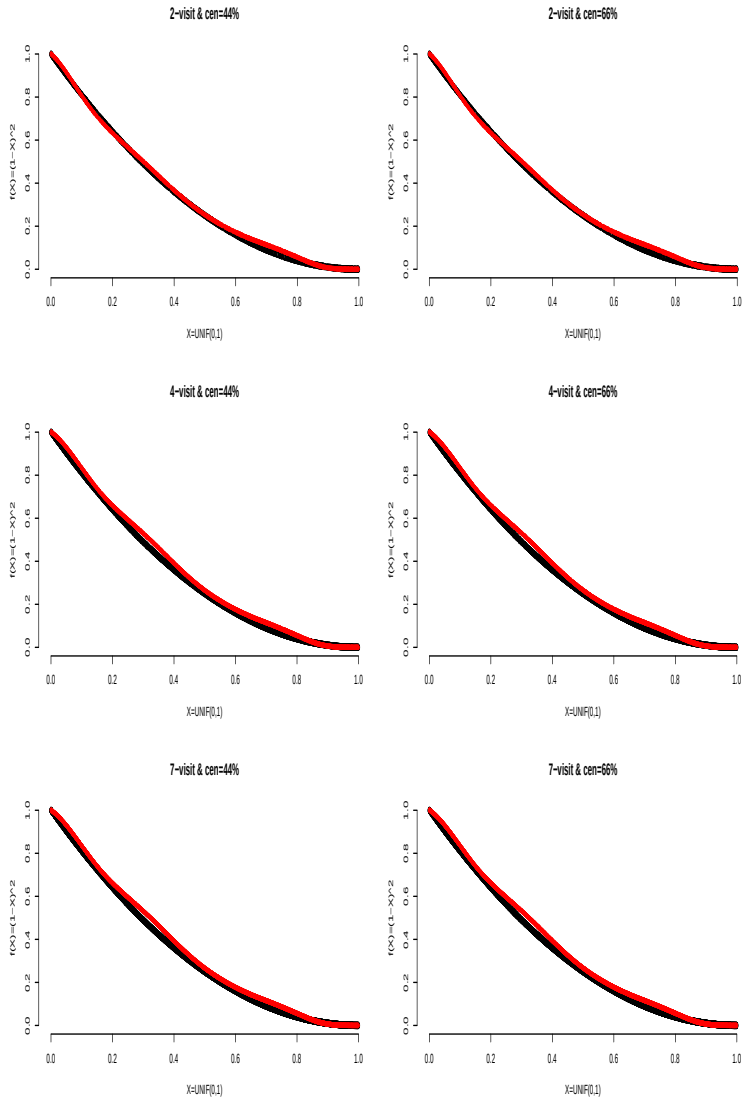
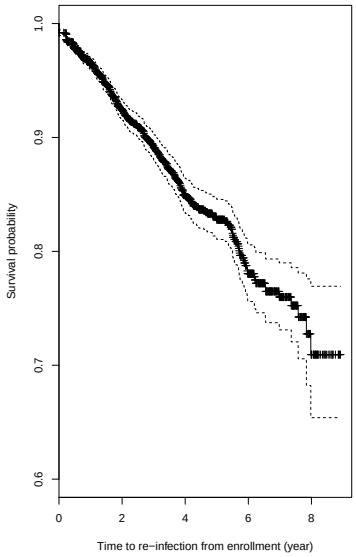
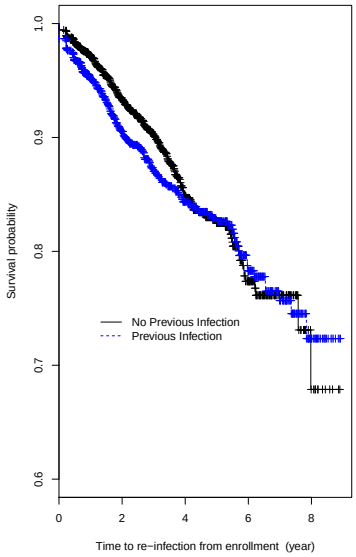


Figure 4.2: Simulation study. Estimated curves from the Cox frailty model; open and black circle represents for true function and closed and red circle represents the estimated function.



(a) Overall KM estimates



(b) KM estimates by past infection status

Figure 4.3: Infection free probability of *Chlamydia trachomatis*

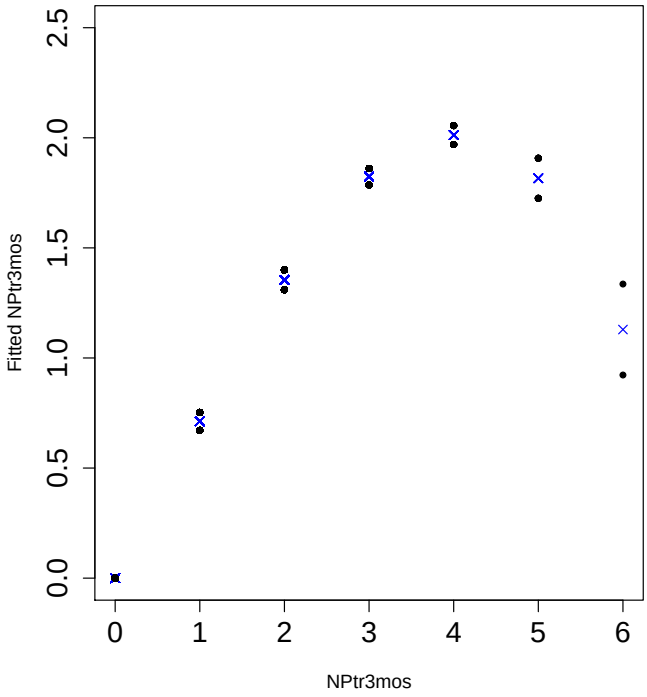


Figure 4.4: Estimate effect of cumulative number of partner in past 3 months with a 95% confidence band ($\times$ represents for estimated values) in the Cox frailty model

Chapter 5

CONCLUSION

Model selection is an important topic in scientific investigation. It not only serves the purpose of dimension reduction, but also ensures the validity of statistical inference. Misspecification of the model could lead to erroneous inferences. Model misspecification usually comes in three different forms: (1) including irrelevant variables, (2) exclusion of relevant variables, and (3) modeling variables in the wrong functional form. The issue is particularly relevant in complex models because complicated modeling structure has made it more difficult for analysts to discern the effects of individual variables for the inclusion/exclusion decision.

My dissertation discusses these issues in cure rate and related models. Specifically, it describes model selection in a class of survival models for situations where a fraction of the subjects are long-term survivors. The work provides a set of model selection tools that help analysts to determine the composition of cure rate or related models. In addition to variable inclusion and exclusion, it presents a method to determine the presence of nonlinear effects. The feature is particularly useful for biomedical investigations where nonlinear relationships are prevalent. My research demonstrates the general applicability of regularization methods. The work shows that statistical concepts such as LASSO and adaptive LASSO could be modified to achieve good model selection results, even in complex statistical models.

What remains to be studied is post selection inference. Although the topic is beyond the scope of the current dissertation, how to approach the issue requires careful and perhaps new thinking about model-based inference. This dissertation has made a tentative step towards that direction, by providing some initial empirical evidence

on the performance of a two-step procedure. An in-depth theoretical examination of the issue appears to be a logical next step.

BIBLIOGRAPHY

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Boag, J. W (1979). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, 11, 15–53.

Berkson, J. and R. P. Gage (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.

Farewell, V. T (1977). A model for a binary variable with time-cesnored observations. *Biometrika*, 64, 43–46.

Farewell, V. T (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38, 1041–1046.

Kuk, A. C. and C. H. Chen (1992). A mixture model combining with logistic regression with proportional hazards regression. *Biometika*, 79(3), 531–541.

Peng, Y and K. B. G. Dear (2000). A nonparametric mixture model for cure rate estimation. Biometrics,56, 237—243.

Sy, J. P. and J. M. Taylor (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56, 227–236.

Lambert, P. C., P. W. Dickman, C. L. Weston, and J. R. Thompson (2010). Estimating the cure fraction in population based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society, Series A*, 59, 35–55.

Chen, Y. Q. and M. C Wang (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association*, 95, 608–618.

Zhang, J. and Y. Peng (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15, 455–467.

Yakovlev, A. Y., A. D. Tsodikov and B Asselain (1996). *Stochastic models of tumor latency and their biostatistics applications*. World Scientific, Singapore.

Tsodikov, D (1998). A proportional hazards model taking account of long term survivors. *Biometrics*, 54, 1508–1516.

Chen, M. H, J. G Ibrahim, and D. Sinha (1999). A Bayesian approach to survival data with a cure fraction. *Journal of the American Statistical Association*, 94, 909–919.

Chen, M. H and J. G Ibrahim (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, 57, 43–52.

Yin, G and JG Ibrahim (2005). Cure rate models: A unified approach. *Canadian Journal of Statistics*, 57, 559–570.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24, 2350–2383.

Tibshirani, R (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 56, 267–288.

Tibshirani, R (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*,16, 385–395.

Fan, J. and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1), 74–99.

Zhang, H. H. and W. Lu (2007). Adaptive LASSO for Cox's proportional hazard model. *Biometrika*, 94, 691–703.

Lin, Y and H. H. Zhang (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34, 2272–2297.

Li, R. and H. Liang (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics*, 36(1), 261–286.

Zhang, H. H, G. Cheng, and Y. Liu (2011). Linear or Nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association*,10, 1099–1012.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*,9, 363–379.

Eilers, P. and B. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*,11(2), 89–121.

Wand, M. P. and J. Ormerod (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian and New Zealand Journal of Statistics*, 50, 179–198.

Lin, C. and S. Halabi (2013). On Model specification and selection of the Cox proportional hazards model. *Statistics in Medicine*, 32, 4609–4623.

Yau, K. K. W. and A. S. K. Ng (2001). Long-term survivor mixture model with random effects: application to a multi-centre clinical trail of carcinoma. *Statistics in Medicine*, 20, 1591–1607.

Yu, B (2008). A frailty mixture cure model with application to hospital readmission data. *Biometrical Journal*, 20(3), 386–394.

Lai, X. and K. W. Yau. (2008). Long-term survivor model with bivariate random effects: application to bone marrow transplant and carcinoma study data. *Statistics in Medicine*, 27, 5692–5708.

Li, Y., E. P. Wileyto, and D. F. Heitjan (2010). Modeling smoking cessation data with alternating states and a cure fraction using frailty models. *Statistics in Medicine*, 29, 627–638.

Rondeau, V. (2010). Statistical models for recurrent events and death: application to cancer events. *Mathematical and Computer modelling*, 52, 949–955.

Rondeau, V., Emmanuel S., Fabien C., Juan G, and Simon M. P. (2011). Cure frailty models for survival data: application to recurrences for breast cancer and to hospital readmission for colorectal cancer. *Statistical Methods in Medical Research*, 0(0), 1–18.

Yan, J. and J. Huang (2012). Model selection for Cox models with time-varying coefficients. *Biometrics*, 68(2), 419–428.

Tepper, RS, CJ Llapur, MH Jones, C. Tiller, C. Coates C, *et al.* (2008). Expired nitric oxide and airway reactivity in infants at risk for asthma. *American Academy of Allergy, Asthma, and Immunology*, 122, 760–765.

Chen, T. (2013). Statistical issues and challenges in immuno-oncology. *Journal for Immunotherapy of Cancer*, 11, 1–18.

Taylor, J. M (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 51, 899–907.

Yakovlev, A. Y., B. Asselain, V. J. Bardou, A. Fourquet, *et al.* (1993). A nonparametric mixture model for cure rate estimation. *Biometrie et Analyse de Dormees Spatio-Temporelles*, 12, 66–82.

Chen, M. H, J. G Ibrahim, and S. R Lipsitz (2002). Bayesian methods for missing covariates in cure rate models. *Lifetime Data Analysis*, 8, 117-146.

Ibrahim, J G and M.H Chen, and D. Sinha (2002). *Bayesian survival analysis*. New York: Springer.

Tsodikov, A. D., J. G Ibrahim, and A. Y. Yakovlev (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Lifetime Data Analysis*, 98, 1063–1078.

Broët, P, Y. De Rycke, P. Tubert-Bitter, J. Lellouch, *et al.* (2001). A semiparametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics*, 57, 844–852.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

He, Z., W. Tu, S. Wang, H. Fu, and Z. Yu (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics*,71, 178–187.

Liu, X., Y Peng, D. Tu, and H Liang (2012). Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in Medicine*, 31, 2882–2891.

Klein, J. P. (1982). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795–806.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression.*The Annals of Statistics*, 12, 758–765.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*,19, 461–464.

Zou, H and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36,1509–1533.

Zhang, Y., R. Li, and C. L. Tsai. (2010). Regularization parameters selection via generalized information criterion. *Journal of the American Statistical Association*, 105, 312–323.

Moore, D and G.P McCabe (2009). *Introduction to the practice of statistics*. Freeman and Company: New York.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802–837.

Corbiére, F., D. Commenges, J. M. Taylor, and P. Joly (2009). A penalized likelihood approach in mixture cure models. *Statistics in Medicine*, 28, 510–524.

Hastie, T. and R. Tibshirani (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46, 1005–1016.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

Wang, L., P. Du, and H. Liang (2012). Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68 (3), 726–735 .

O'Sullivan, F (1986). A statistical perspective on ill posed inverse problems.*Statistical Science*,1, 505–527.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Gray, R. (1992). Methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87, 942–951.

Wang, Y (2011). *Smoothing Splines: Method and Applications*. Boca Raton: CRC Press.

P̃ena, J. M. (1997). B spline and optimal stability. *Mathematics of Computation*, 66(220), 1555–1560.

Tu, W., G. J. Eckert, L. DiMeglio, Z. Yu, *et al.* (2011). Intesnsified effects of adiposity on blood pressure in overweight and obese children. *Hypertension*, 58, 818–824.

National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents (2004). *Pediatrics*, 114, 555–576.

Li, Z., H. Liu, and W. Tu (2015). A sexually transmitted infection screening algorithm based on semiparametric regression models. *Statistics in Medicine*, 34(20), 2844–2857.

Kelly, P. J. and L. Lim (2000). Survival analysis for recurrent event data: an application to childhood infection diseases. *Statistics in Medicine*, 19, 13–33.

Scolas, S, Ghouch A, Legrand C, and Oulhaj A (2016). Variable selection in flexible parametric mixture cure model with interval-censored data. *Statistics in Medicine*, 35(7):1210-1225.

Stone, CJ (1985). Additive regression and other nonparametric models. *The Annals of statistics*, 13(2): 689–705.

Peng, Y. and JM Taylor (2014). Cure models. *Handbook of survival analysis* Boca Raton: CRC Press.

R Development Core Team (2014). *R: A language and environment for statistical computing*.Vienna, Austria: R Foundation for Statistical Computing.

Tu, W., B. E. Batteiger, S Wiehe, S. Ofner, B. Van Der Pol, *et al.*(2009). Time from first intercourse to first sexually transmitted infection diagnosis among adolescent women. *Archives of pediatrics & adolescent medicine*, 163(12): 1106–1111.

Batteiger, B.E., W. Tu, S. Ofner, B. Van Der Pol, *et al.* (2010). Repeated *Chlamydia trachomatis* genital infections in adolescent women. *Journal of Infectious Diseases*, 201(1): 42–51.

Masud, A., W. Tu, and Z. Yu (2016) . Variable selection for mixture and promotion time cure rate models. *Statistical Methods in Medical Research*, 0(0), 1-15.

Yu, Z., X. Lin, and W. Tu (2012). Semiparametric frailty models for clustered failure time data. *Biometrics*, 68(2): 429–436.

CURRICULUM VITAE

Abdullah Al Masud

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, Indiana, 2017 (minor in Epidemiology)
- M.S. in Statistics and Financial Economics, Utah State University, Logan, Utah, 2012
- B.S. in Statistics, University of Dhaka, Dhaka, Bangladesh, 2008

WORKING EXPERIENCE

- Senior Statistician-Oncology, AbbVie, North Chicago, Illinois, January 2017–Present
- Summer Intern:
  - Bristol-Myers Squibb, Wallingford, Connecticut, May 2016– August 2016
  - Pfizer Inc, New York City, New York, May 2015– August 2015
  - Pinnacle Solutions Inc, Indianapolis, Indiana, May 2014– August 2014
- Research Assistant, Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, August 2015–December 2016
- Teaching Assistant:
  - Department of Mathematics, Indiana University Purdue University Indianapolis, Indianapolis, Indiana, Augusts 2013–May 2015
  - Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, Augusts 2012–July 2013
  - Department of Mathematics & Statistics, Utah State University, Logan, Utah, Augusts 2009–July 2011

SELECT PUBLICATIONS

- Masud, A., Tu, W., and Yu, Z. (2016). Variable Selection for Mixture and Promotion Time Cure Rate Models. *Statistical Methods in Medical Research*, **0(0)**: 1–15.
- Kwag, A. and Masud, A. (2013). Modeling and Predicting Stock Returns: The Rule of Parsimony. *POSRI Business Economic Research Article*, **13(2)**: 149–187.
- Chowdhury, Z., Campanella, L, Gray, C., Masud, A., Pennise, D., and Zhuzhang, X (2012). Evaluation and Modeling of Indoor Air Pollution in Rural Households with Multiple Stove Interventions in Yunnan, China. *Atmospheric Environment*, **67**: 161–169.
- Chowdhury, Z., Leah, T. L., Karen, C., Masud, A., Alauddin, M., Hossain, M., Zakaria, ABM, and Hopke, P. H (2012). Quantification of Indoor Air Pollution from Using Cook Stoves and Estimating its Health Effects in Northwest Bangladesh. *Aerosol and Air Quality Research*, **12**: 463–475.

PRESENTATIONS

- Assessment of Weighted Log-Rank Test for Immuno-Oncology Trials, Global Biometric Sciences (GBS)-Oncology at Bristol-Myers Squibb, August 2016. Wallingford, CT.
- Improved Finkelstein-Schoenfeld (FS) Test in Clinical Trial applications. Oral Presentation, Global Innovative Pharma Business (GIPB) at Pfizer Inc, August 2015. Manhattan, NY.
- Variable Selection for Mixture And Promotion Time Cure Rate Models. Oral Presentation, International Chinese Statistical Association (ICSA) Applied Statistics Symposium, June 2016. Atlanta, GA.

- A Comparison of Weighted P-Values And Multi-Stage Analyses in Multiple Hypothesis Testing. Poster, Applied Statistics in Agriculture, April 2010. Manhattan, KS.